Mixed Integer Linear Programming for Traffic Signal Control

Iain Guilliard

A thesis submitted for the degree of
Doctor of Philosophy of
The Australian National University

September 2020

Declaration

I hereby declare that this thesis is my original work which has been done in collaboration with other researchers. This document has not been submitted for any other degree or award in any other university or educational institution. Parts of this thesis have been published in collaboration with other researchers in international conferences and journals as listed below:

- **(Chapter 3)** Guilliard, I.; Sanner, S.; Trevizan, F. W.; and Williams, B. C., 2016. Nonhomogeneous time mixed integer linear programming formulation for traffic signal control. *Transportation Research Record: Journal of the Transportation Research Board*, 2595 (2016), 128–138. doi: 10.3141/2595-14.
- **(Chapter 4)** Guilliard, I.; Trevizan, F. W.; and Sanner, S., 2020. Mitigating the impact of light rail on urban traffic networks using mixed integer linear programming. *IET Intelligent Transport Systems*, 14, 6 (2020), 523–533. doi: 10.1049/2019.0277. URL https://doi.org/10.1049/iet-its.2019.0277.

Iain Guilliard
3 September 2020

Acknowledgments

I would like to thank Dr Felipe Trevizan and Dr Scott Sanner for their supervision of this thesis. Felipe, for his endless patience, pragmatism and advice, and Scott, for his motivation and critical feedback, but also for his encouragement to pursue this research in the first instance, and the opportunities that it created. And also to the chair, Assoc. Prof. Patrik Haslum, for his help and guidance throughout the process.

To Greta, for her unwavering support, patience and friendship.

Abstract

As urban traffic congestion is on the increase worldwide, it is critical to maximize capacity and throughput of existing road infrastructure through optimized traffic signal control. This thesis builds on the body of work using mixed integer linear programming (MILP) approaches that attempt to jointly optimize traffic signal control over an entire traffic network and specifically improve the scalability of these methods to larger numbers of intersections.

The main contribution of this thesis is the Queue Transmission Model (QTM), a MILP formulation for traffic signal control, which can be evaluated at non-homogeneous time intervals. This property can be used to extend the planning horizon of a traffic signal controller by strategically adjusting the spacing between samples. By using more samples near the start, and fewer at later stages, the planner can adapt to long term changes in traffic flow, while improving the short-term fidelity. The performance of this approach is evaluated on several networks of differing topology, and the results show that, compared to homogeneous time control with the same number of intervals, this method is able to produce solutions with substantially lower overall travel times, and better per vehicle delay distribution.

Another contribution of this thesis is modeling light rail systems that share intersections with vehicle traffic, to aid the many cities considering light rail in understanding its impact on signal timing and delay. A method is described to incorporate light rail schedules and fixed time control as additional constraints on QTM signal timing, such that the controller is able to produce signal plans that take the light rail into account when optimizing for the vehicle traffic. A micro simulator is then used to evaluate the performance of all these extensions, comparing fixed time control to optimized adaptive control, using multiple scenarios of network topology, light rail schedules and traffic levels. The results show that optimized adaptive control plans have substantially lower average delay, better per vehicle delay distribution and lower numbers of stops. In some scenarios, switching to optimized adaptive control nullifies the impact on signal timing of introducing light rail, while persisting with fixed time control requires a significant reduction in traffic levels to achieve the same outcome.

The final contribution of this thesis is to compare QTM with several different formulations for MILP based traffic signal optimization both theoretically and empirically. First, it is demonstrated using variational theory, that all the models find equivalent discrete solutions to kinematic wave theory. This result is used to finally address the issue of unintended vehicle withholding, and to show for the first time that withholding has no impact on the optimality of the solutions. Finally, a series of experiments is run on networks of increasing size, using vehicle platoons of varying length and arrival time. The results show that when comparing both the solve time and the quality, QTM is able to find better policies with lower delay, and in less time than the other formulations.

Contents

Symbols

ρ^{jam}	maximum density of a section of road way
q^{c}	capcity flow of a section of road way
ρ^{c}	desity at the capcity flow q^{c}
$\mathcal{Q}_i^{\mathcal{P}}$	set of traffic light phases controlling the outflow of queue i
$\mathcal{P}_\ell$	set of phases of ℓ
$q_{i,n}$	$\in [0, \text{Q}_i]$ traffic volume waiting in the stop line of queue i at the beginning of interval n
$f_{i,n}^{\text{in}}$	$\in [0, I_{i,n}]$ inflow to the network via queue i during interval n
$f_{i,n}^{\text{out}}$	$\in [0, \text{F}_i^{\text{out}}]$ outflow from the network via queue i during interval n
$f_{i,j,n}$	$\in [0, \text{F}_{i,j}]$ flow from queue i into queue j during interval n
$p_{\ell,k,n}$	$\in \{0, 1\}$ phase activation parameter during interval n for phase k of ℓ
$d_{\ell,k,n}$	$\in [0, \Phi_{\ell,k}^{\max}]$ phase duration parameter during interval n for phase k of ℓ
Q_i	maximum capacity of queue i
F_i^{out}	maximum traffic flow from i to the outside of the modeled network
T_i^{p}	time required to traverse link i and reach the stop line
$\text{F}_{i,j}$	maximum flow from link i to link j
$\text{Pr}_{i,j}$	turn probability from link i to link j
$\Phi_{\ell,k}^{\max}$	maximum allowed time for phase $k \in \mathcal{P}_\ell$
$\Phi_{\ell,k}^{\min}$	minimum allowed time for phase $k \in \mathcal{P}_\ell$
$\Psi_\ell^{\max}$	maximum allowed cycle time for light ℓ
$\Psi_\ell^{\min}$	minimum allowed cycle time for light ℓ
k	density in vehicles per meter along a section of road way
q	flow in vehicles per second along a section of road way
$Q(k)$	fundamental relationship between flow and density
$N(x, t)$	cumulative vehicle function at x, t

Chapter 1

Introduction

1.1 Motivation

Beyond the basic requirements of safe and orderly traffic coordination, traffic signal control schemes have long sought to improve utilization. Existing infrastructure makes widespread use of fixed time control, optimized offline for average conditions, but often with some form of real time actuation to improve responsiveness. Several adaptive control methods have found widespread adoption [Sims and Dobinson, 1980; Hunt et al., 1981], where a baseline fixed schedule is adapted to sensed traffic conditions using heuristically driven feedback control algorithms.

Optimized control strategies promise to further improve utilization. Typically such approaches use some form of *model predictive control* [Garcia et al., 1989], where the control input (the signal plan) is optimized based on the predicted response of a model of the system dynamics (the traffic flow), given an objective (to minimize delay). We can separate optimal strategies broadly into two categories: approximate and exact. Approximate methods use heuristics and decomposition to reduce the computational load, at the sacrifice of true optimality. Exact methods on the other hand, use mathematical programming to find the *exact* optimal solution. Of particular interest are those methods employing *Mixed Integer Linear Programming* (MILP). Among the first works to describe a MILP for traffic signal optimization is that of Garter [Gartner et al., 1974]. Gartner proposes using a platoon based model of traffic flow to optimize the basic traffic signal parameters, with the emphasis on finding coordinated strategies along arterial routes. The Cell Transmission Model (CTM) [Daganzo, 1994] provided a practical framework to solve kinematic wave (KW) based macroscopic traffic flows described by the "LWR" theory of Lighthill, Whitham and Richards [Lighthill and Whitham, 1955; Richards, 1956]. Lo [Lo, 1998] was the first to propose a MILP using CTM, capturing the queuing dynamics of signalized traffic flow, and offering *globally* optimal solutions spanning multiple intersections. This was followed by MILP formulations for the LWR based Link Transmission Model (LTM) [Yperman et al., 2005; Hajiahmadi et al., 2012], and for the Link based Kinematic Wave Model (LKWM) [Han et al., 2012]. Using the vatiational calculus of Hamiltonian-Jacobi theory, Daganzo developed Variational Theory (VT) [Daganzo, 2005a,b], as a general framework for solving KW based traffic problems, and later [Wada et al., 2017] cast VT as a MILP for traffic signal optimization.

1.2 Contributions

The major contribution of this thesis is the queue transmission model (QTM), a non-homogeneous time MILP-based model of joint intersection control that blends elements of cell-based and link-based modeling approaches. The QTM offers the following key benefits:

- Unlike previous CTM-based joint intersection signal optimization [Lo, 1998; Lin and Wang, 2004], the QTM is intended for *non-homogeneous* time steps that can be used for control over large horizons.
- Any length of roadway without merges or diverges can be modeled as a single queue leading to compact QTM MILP encodings of large traffic networks (i.e., large numbers of cells and their associated MILP variables are not required *between* intersections).

Further, the free flow travel time of a link can be modeled exactly, independent of the discretization time step, while CTM requires a further increased discretization to approach the same resolution.

- The QTM accurately models fixed travel time delays critical to green wave coordination as in Gartner, Gartner and Stamatiadis, and He [Gartner et al., 1974; Gartner and Stamatiadis, 2002; He et al., 2011] through the use of a non-first order Markovian update model and further combines this with fully joint intersection signal optimization in the spirit of Lo, Lin and Wang, and Han [Lo, 1998; Lin and Wang, 2004; Han et al., 2012].

Experiments with this novel QTM-based MILP control in a range of traffic networks, demonstrate that the non-homogeneous MILP formulation achieves (i) substantially lower delay solutions, (ii) improved per-vehicle delay distributions, and (iii) more optimal travel times over a longer horizon in comparison to the homogeneous MILP formulation with the same number of binary and continuous variables.

One benefit of optimiztion based control is the that additional contraints can be added to the model without having to reformulate. Using this approach:

1. QTM is leveraged with the addition of light rail constraints for optimal adaptive control, by actively optimizing the signal plan over a fixed time horizon from the current traffic state.

2. A novel method that extends QTM to optimize conventional fixed-time signal control schedules, with and without light rail schedule constraints and common cycle length constraints, which can be incorporated *immediately* into existing fixed-time traffic controller infrastructure,

3. QTM's support for multiple phases per cycle is utilized as a novel way to model the lost time associated with signal changes, and show that it is critical to finding optimized signal plans.

4. A comprehensive suite of experiments using a microsimulator to validate the effectiveness of these contributions, both quantitatively and through visual inspection of the simulation results.

Fixed and optimized control is compared on several traffic networks, with and without light rail, to show that optimal adaptive control can reduce traffic delay by up to 57.8% over optimal fixed-time control when light rail is introduced, and virtually nullifies its impact when compared to using fixed-time control before the introduction of light rail.

A key problem faced by MILP based traffic flow models making linear relaxations is lost time: traffic may not advance downstream even when there is free capacity to do so, requiring additional binary variables to directly model the non-linearities, or additional objective terms. Using Variation Theory, it is shown for the first time that traffic withholding has no impact on the optimal solution. Additionally, Variational Theory is extended to include turn probabilities which shows that solving a QTM traffic signal optimization with turn probabilities is equivalent to a stochastic shortest path problem in space-time.

To date there has been no comprehensive comparison of the various MILP based approaches to traffic signal optimization. This work contributes the first extensive comparison of CTM, LTM, VT and QTM with micro-simulation validation.

1.3 Thesis Outline

Chapter 2 introduces basic concepts of traffic engineering needed to understand the algorithms described later in this thesis. LWR theory as a macroscopic model of traffic flow is introduced and several discrete models are derived from the theory. These models are then cast as mixed integer linear programs for traffic signal control.

Chapter 3 formally introduces the Queue Transmission Model as a model of traffic flow with non-homogeneous time steps and shows how to encode it as a linear program for computing traffic flows. Next, the traffic signals states are represented by discrete phase variables that are optimized subject to a delay minimizing objective and standard minimum and maximum time constraints for cycles and phases. This results in the MILP formulation of traffic signal control. Finally a series of experiments using QTM-based MILP control in a range of traffic networks is used to demonstrate that the non-homogeneous MILP formulation achieves (i) substantially lower delay solutions, (ii) improved per-vehicle delay distributions, and (iii) more optimal travel times over a longer horizon in comparison to the homogeneous MILP formulation with the same number of binary and continuous variables.

Chapter 4 looks at novel extensions to QTM and provides micro-simulation validation. QTM is leveraged with the addition of light rail constraints, and a novel method that extends QTM to optimize conventional fixed-time signal control schedules with common cycle length constraints. QTM's signal timing constraints are extended to model the lost time associated with signal changes, and it is shown that modeling lost time is critical to finding optimized signal plans. A comprehensive suite of experiments using a microsimulator is used to validate the effectiveness of these contributions, both quantitatively and through visual inspection of the simulation results.

Chapter 5 looks at how QTM relates to the LWR kinematic wave equation using Variational Theory and compares QTM with CTM, LTM and VT models. QTM is extended to fully model multiple shockwaves and the equivalence between QTM, LTM, CTM, and VT is formally derived using Variational Theory, which provides a framework for understanding the traffic withholding problem. For the first time it is shown that traffic withholding has no impact on the optimal solution. QTM is then used to extend VT with turn probabilities, which shows, using Variational Theory, that QTM is equivalent to a stochastic shortest path problem, and solves a Markov Decision Problem value function. A novel continuous time MILP for exactly solving QTM is then described.

Finally, in Chapter 6, using the new framework established in the previous chapters, QTM performance is compared to CTM, LTM and VT in terms of both solve time and solution quality.

Chapter 2

Background and Related Work

2.1 Overview

This chapter will introduce the basic concepts of traffic theory referred to throughout this work, and describe several important traffic flow models and their Mixed Integer Linear Programming (MILP) formulations.

Following an over view of basic traffic flow concepts in Section 2.2, Section 2.3 outlines LWR theory, a first order macroscopic model of traffic flow that underpins all the methods described herein. Section 2.4 describes one of the first successful discrete time models based on LWR theory, the Cell Transmission Model (CTM). Newell's Theory for solving LWR problems using cumulative counts is introduced in Section 2.5, and the Link Transmission Model (LTM), a discrete time formulation of Newell's Theory, is outlined in Section 2.6. Variational Theory is introduced in Section 2.7, as a further extension of Newell's theory for solving general non-homogeneous LWR problems. And finally, in Section 2.8, MILP formulations of all these models for traffic signal optimization are detailed in full.

2.2 Traffic Flow Modelling

This section will introduce some basic concepts in traffic flow modelling and analysis, used throughout this thesis.

2.2.1 Space-Time Diagrams

One of the most effective ways to visualize the evolution of traffic flowing along a roadway is through the space-time diagram. The trajectory of each vehicle is plotted on a graph with the time of travel along the x-axis and position on the road along the y-axis. See Fig. 2.1. Several useful parameters can be ascertained from the space-time diagram:

- velocity of a vehicle, or indeed other phenomena moving with the traffic such as shockwaves, can be determined from the slope of its trajectory.
- traffic density in vehicles per unit distance is indicated by the vertical spacing of the traces.
- traffic flow in vehicles per unit time is indicated by horizontal spacing of the traces.
- crossing traces indicate overtaking traffic

In general, the density, k, and flow, q, can be calculated for any arbitrary region on the space-time diagram with the equations (2.1) and (2.2), where A is the area of the region, and t_i and x_i are respectively the travel time and distance travelled through the region by vehicle i.

$$k = \frac{\sum_i t_i}{A} \tag{2.1}$$

$$q = \frac{\sum_i x_i}{A} \tag{2.2}$$

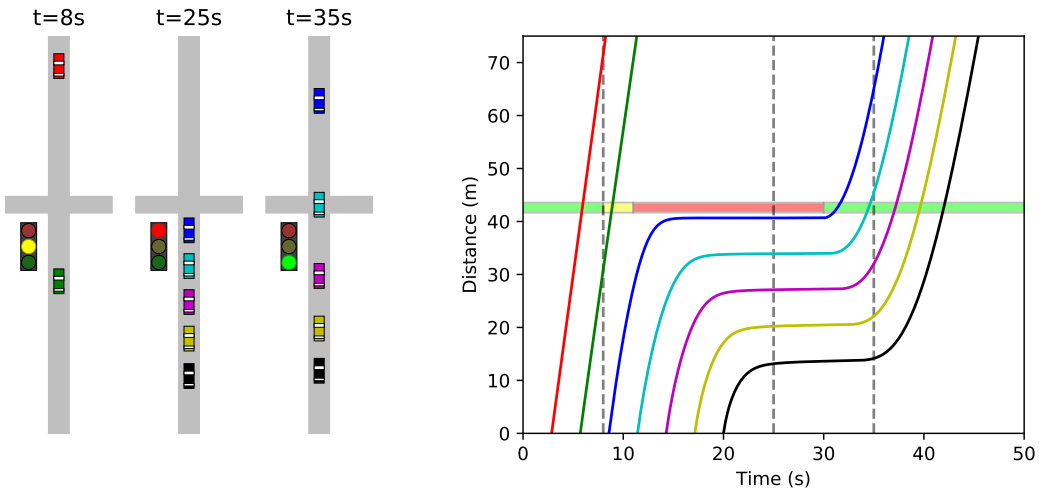


Figure 2.1: Space-Time diagram of vehicle flow along a roadway. The first two vehicles (red and green) traverse the length of the roadway unimpeded (t=8s plot). The slope of their trajectories is the free flow speed. A red traffic signal change at 10s brings the next five vehicles to a standstill, forming a stop-line queue (t=25s plot). The queue is indicated by horizontal trajectories until the return of a green signal at 30s. From t=35s the 5 vehicles are de-queuing and the trajectories return to the free flow speed. Backwards waves during the queuing and de-queuing process can been seen as inflections travelling through the trajectories.

2.2.2 Cumulative Curves

If a counter is placed on the road side that senses and keeps a count of passing vehicles, then a cumulative count is generated over time. For example, see Fig. 2.2. While the cumulative count is a step function, typically with large traffic volumes it is interpreted as piece-wise smooth function referred to as a cumulative curve, N (t), where the floor, $\lfloor N(t) \rfloor$ is the cumulative vehicle count at time t. Cumulative curves have some useful properties in traffic engineering. Consider two cumulative curves generated by sensors separated by some distance d along the road at x_1 and x_2, then

- the horzontal distance between two curves at level n gives the travel time between the two points for vehicle n.
- the vertical distance between the two curves at time t gives the number of vehicles currently travelling between points x_1 and x_2.
- the slope of the curve gives the flow rate in vehicles per second.

We will see that the properties of cumulative vehicle functions will underpin several of the traffic models descried later in this chapter.

2.2.3 Fundamental Relationship of Flow and Density

If we sample regions over a space-time plot of traffic flow, and calculate the density and flow for the regions, we can plot a curve relating flow as a function of density, often referred to as

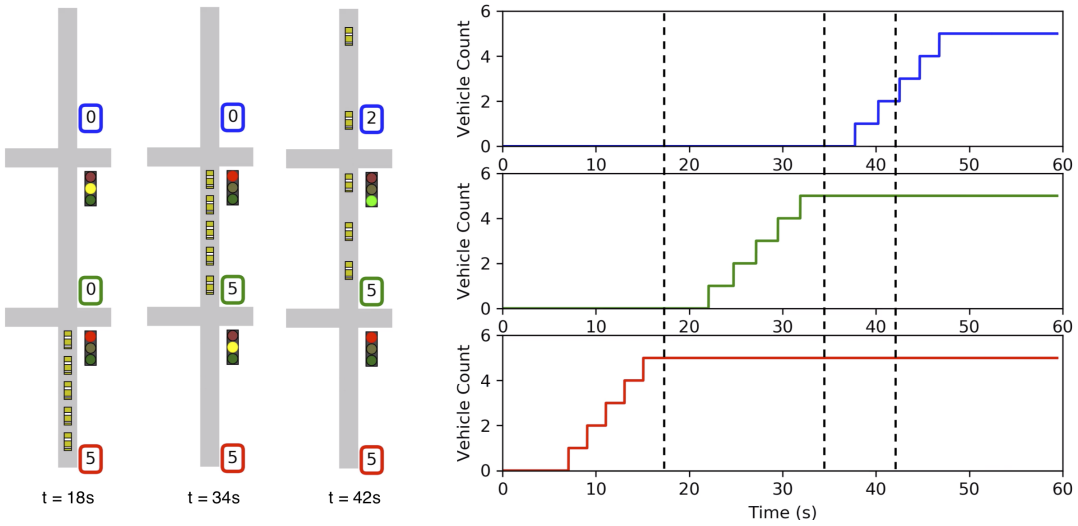


Figure 2.2: Cumulative vehicle counts along a roadway at distances of 0 m, 45 m and 70 m. The state of the cars and counters shown on the left plots is indicated by the vertical dotted lines in the curve plots at t = 18s, t = 34s, and t = 42s, respectively

the fundamental diagram of traffic. The relationship is typically concave and has been well observed in many different traffic flow scenarios [Kühne and Gartner, 2011]. As a function it is assumed to be piece-wise differentiable, non-negative, and defined over an interval $[0, \rho^{\text{jam}}]$. At a critical density, ρ^{c}, the maximum flow, q^{c} is observed. See Fig. 2.3 for an example of the relationship between a fundamental diagram and a space time diagram of traffic flow.

2.2.4 Newell Surfaces

If a family of cumulative curves is generated along a road at every point, and the curves are then plotted as a surface above a space time diagram, they would form a Newell surface. The Newell function $N(x,t)$ gives the cumulative count value on the surface at location (x,t) on the space-time diagram. See Fig. 2.4 for an example of a Newell surface. Newell surfaces have some interesting properties. The first partial derivatives of $N(x,t)$ are the flow and density functions given by (2.3) and (2.4) respectively, and the second partial derivatives give the conservation (2.5). However it should be noted that these derivatives hold everywhere except possibly along certain curves in the surface known as shocks, that will be introduced in Section 2.3.1

$$q(x,t) = \frac{\partial}{\partial t} N(x,t) \tag{2.3}$$

$$k(x,t) = -\frac{\partial}{\partial x} N(x,t) \tag{2.4}$$

$$\frac{\partial^2}{\partial x \partial t} N(x,t) = \frac{\partial^2}{\partial t \partial x} N(x,t) \tag{2.5}$$

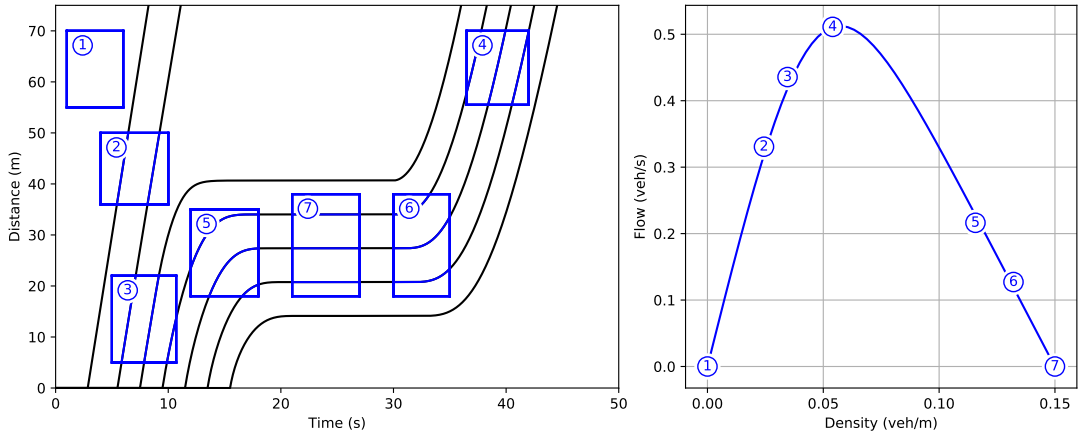


Figure 2.3: Relationship between a space-time plot of traffic flow and the fundamental diagram. In each of the numbered regions in the left plot, the flow and density is calculated and plotted as a numbered point in the right plot, and each corresponds to a valid flow-density pair on the fundamental diagram.

2.3 Lighthill-Witham-Richards Model

The Lighthill-Witham-Richards model is a macroscopic model that relates traffic flow to a hydro dynamic flow was first proposed by Lighthill and Witham [Lighthill and Whitham, 1955] and independently by Richards [Richards, 1956], and is often referred to in the literature as "LWR theory". The classic kinematic wave equation (2.6) describes a conservation law between changes in flow and density.

$$\frac{\partial}{\partial t}k(x,t) + \frac{\partial}{\partial x}q(x,t) = 0 \tag{2.6}$$

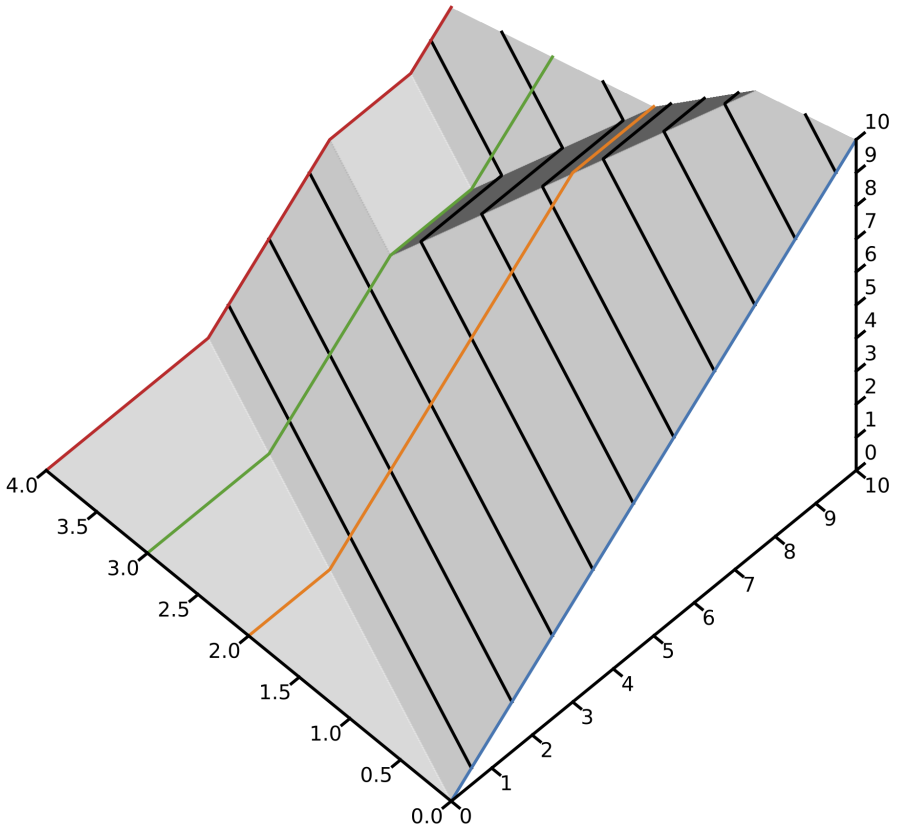
With the addition of function relating density, k, and the flow, q given by $q(x,t) = Q\left(k(x,t), x, t\right)$, (2.6) takes the form (2.7), where the relation Q is the fundamental diagram of traffic flow.

$$\frac{\partial}{\partial t}k(x,t) + \frac{\partial}{\partial x}Q\left(k(x,t), x, t\right) = 0 \tag{2.7}$$

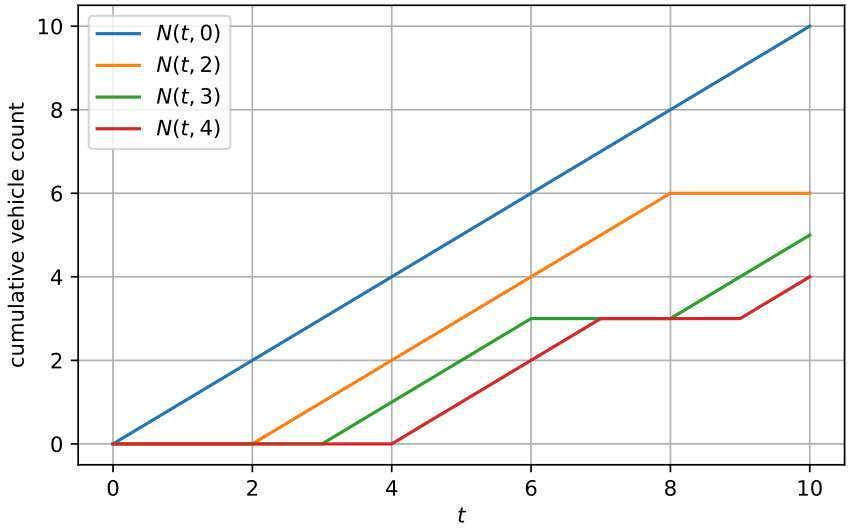
Some examples of LWR solutions with different fundamental diagrams are shown in Fig. 2.5.

2.3.1 Shock Waves

LWR theory predicts the existence of shocks travelling as waves through the traffic, both forwards and backwards, wherever there is a discontinuity in the density. These shock waves can be clearly seen in on a Newell surface, for example in Fig. 2.6.

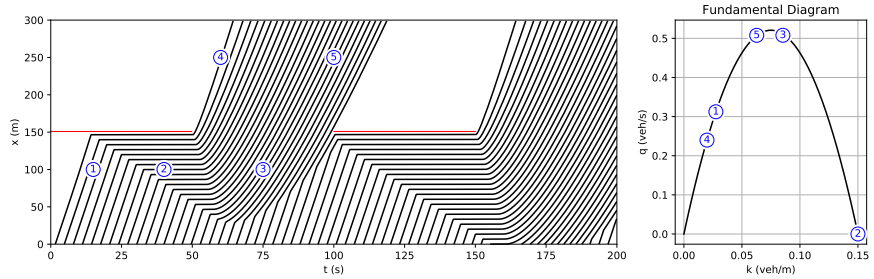


(a)

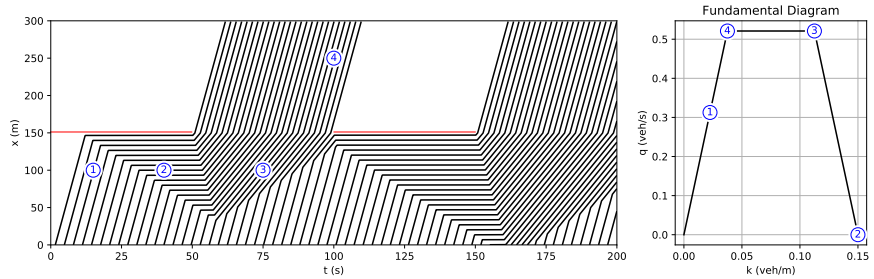


(b)

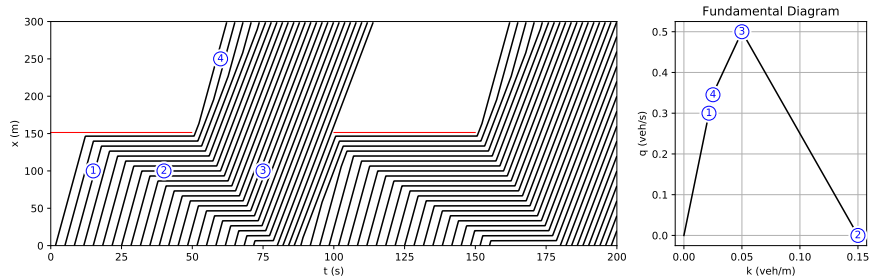
Figure 2.4: (a) An example of a Newell surface formed by the function $N(x,t)$, the cumulative vehicle count at position (x,t) on a space-time diagram. Horizontal contours represent vehicle trajectories, and the vertical slices represent the cumulative curves at $x = 0$, $x = 2$, $x = 3$, and $x = 4$, as shown in (b).



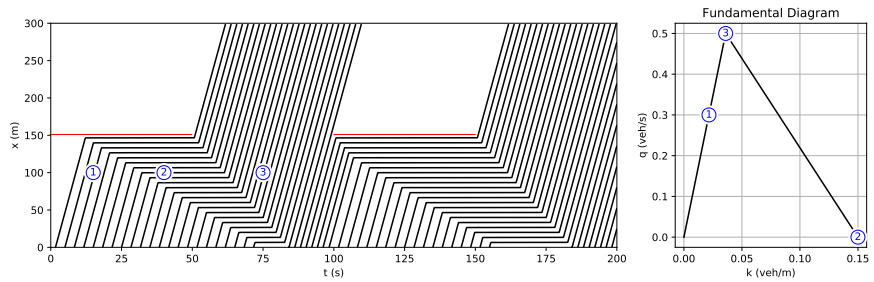
(a)



(b)



(c)



(d)

Figure 2.5: Solutions to the LWR equation for different fundamental diagrams. (a)—(d) show the LWR modeled traffic flow along a 300m section of roadway with a traffic signal at $x = 150$ m. When the signal is red $Q\left(k(x,t),x,t\right) = 0$ at $x = 150$, otherwise $Q\left(k(x,t),x,t\right)$ is given by the fundamental diagram to the left of each plot. Each trace represents the path of a vehicle over time, and the propagation of shockwaves through the traffic is clearly visible. The numbered labels within the flow correspond to points along the fundamental diagram.

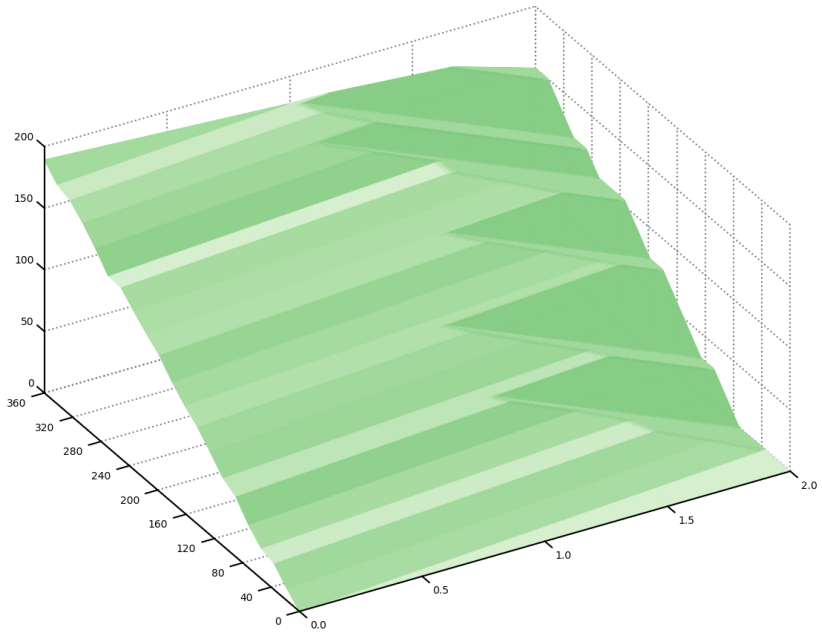


Figure 2.6: Examples of shock waves emanating from a signalized intersection. The plot shows a Newell surface for the region of a roadway immediately upstream of the intersection as it evolves over a period of 360 seconds. The boundary between free flow and congestion can be clearly seen zig-zagging across the surface, with the shock paths travelling backwards and forwards along the link.

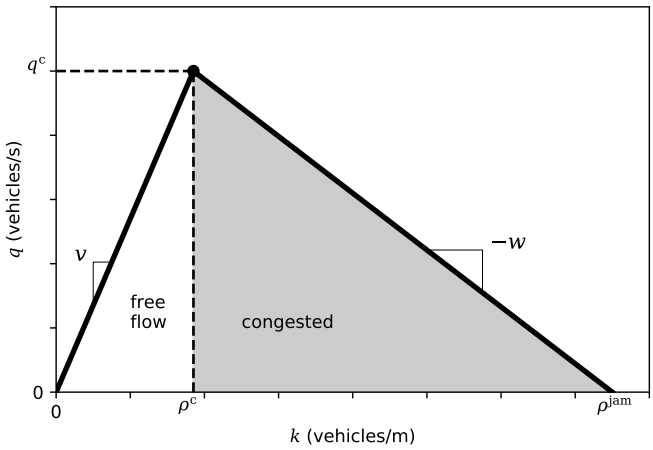


Figure 2.7: Triangular fundamental diagram relating density, k, to flow, q. The maximum capacity flow rate is q^c at a density of ρ^c, and divides the traffic into two regimes: free flow when the density is lower than ρ^c, and congested when the density is higher. The slope of the curve in the free flow region is the forward wave speed, v, and in the congested region is $-w$, where w is the backward wave speed.

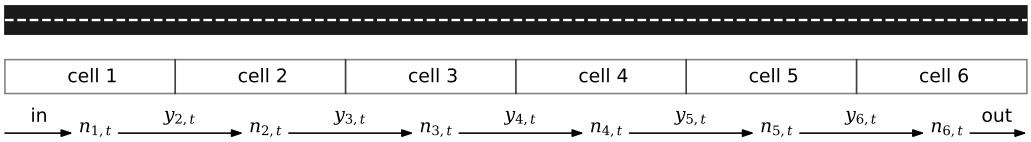


Figure 2.8: The Cell Transmission model. A homogeneous section of roadway is modeled as a chain of cells. At each time step, vehicles are transmitted between cells at flow rates determined by a trapezoidal function (fundamental diagram) of the cell densities, and the state of the cells is updated.

2.3.2 Hamilton-Jacobi Equation

Since the Newell function has the property of the conservation law given by (2.5), a special form of the LWR equation known as the Hamilton-Jacobi Equation, can be derived by substituting the partial derivatives of $N(x,t)$ for flow and density, (2.3) and (2.4), into the fundamental relation $q(x,t) = Q\left(k(x,t)\right)$, giving (2.8).

$$\frac{\partial}{\partial t}N(x,t) - Q\left(-\frac{\partial}{\partial x}N(x,t)\right) = 0 \tag{2.8}$$

The Hamilton-Jacobi form of the LWR equation permits the derivation of several link based models, that will shown in the following sections.

2.4 The Cell Transmission Model (CTM)

The Cell Transmission Model (CTM) of Daganzo [Daganzo, 1994, 1995] is a discrete time solution to LWR theory using finite differences to approximate equation (2.7), with $Q()$ given by a triangular or trapezoidal fundamental diagram (Fig. 2.7). The roadway is divided up into cells of length equal to the distance traveled in one time step at the free flow speed v. A cell is characterized by a maximum flow rate of Q vehicles per time step, and a maximum occupancy of N vehicles. If consecutive cells are indexed $1, 2, 3\ldots$, then the the number of vehicles, $y_{i,t}$, flowing from cell $i-1$ to cell i during the interval $[t, t+\Delta t)$ is given by (2.9),

$$y_{i,t} = \min\left\{Q, \quad n_{i-1,t}, \quad \frac{w}{v}\left(N - n_{i,t}\right)\right\} \tag{2.9}$$

where $n_{i,t}$ is the number of vehicles in cell i at time t, and the ratio $\frac{v}{w}$ represents the backwards wave speed when expressed in units of cell lengths per time step. The state update for cell i at the end of interval $[t, t+\Delta t)$ is given by (2.10).

$$n_{i,t+1} = n_{i,t} + y_{i,t} - y_{i+1,t} \tag{2.10}$$

Equations (2.9) and (2.10) define CTM. To see how CTM approximates the LWR equation (2.7), consider that a trapezoidal fundamental diagram is a concave piece-wise linear function of

$k(x,t)$ defined over the interval $[0, \rho^{\text{jam}}]$, given by (2.11),

$$q(x,t) = \min\left\{q^{\text{c}}, \quad vk(x,t), \quad w(\rho^{\text{jam}} - k(x,t))\right\} \tag{2.11}$$

With a time step of Δt, and the cell length of $\Delta x = v\Delta t$, we can derive the flows and densities:

$$\rho^{\text{jam}} = \frac{N}{\Delta x} = \frac{N}{v\Delta t} \tag{2.12}$$

$$k(x,t) = \frac{n_{i,t}}{\Delta x} = \frac{n_{i,t}}{v\Delta t} \quad \text{or,} \quad k(x,t) = \frac{n_{i-1,t}}{\Delta x} = \frac{n_{i-1,t}}{v\Delta t} \tag{2.13}$$

$$q^{\text{c}} = \frac{Q}{\Delta t} \tag{2.14}$$

$$q(x,t) = \frac{y_{i,t}}{\Delta t} \tag{2.15}$$

Note that for estimating the density at x in (2.13), we can choose either the cell i downstream of x or the cell $i-1$ upstream. Substituting these into (2.11),

$$\frac{y_{i,t}}{\Delta t} = \min\left\{\frac{Q}{\Delta t}, \quad v\frac{n_{i-1,t}}{v\Delta t}, \quad w\left(\frac{N}{v\Delta t} - \frac{n_{i,t}}{v\Delta t}\right)\right\} \tag{2.16}$$

and normalizing the units of time so that $\Delta t = 1$ gives the CTM equation for $y_{i,t}$, (2.9) and connects $y_{i,t}$ to $q(x,t)$:

$$y_{i,t} = \min\left\{Q, \quad n_{i-1,t}, \quad \frac{w}{v}(N - n_{i,t})\right\} = q(x,t) \tag{2.17}$$

As $\Delta x \to 0$ and $\Delta t \to 0$, $\frac{\partial}{\partial x}q(x,t)$ is approximated by $y_{i+1,t} - y_{i,t}$, and $\frac{\partial}{\partial t}k(x,t)$ is approximated by $n_{i,t+1} - n_{i,t}$. Substituting into (2.6) and rearranging (2.18) gives (2.19), the CTM state update equation, (2.10).

$$(y_{i+1,t} - y_{i,t}) + (n_{i,t+1} - n_{i,t}) = 0 \tag{2.18}$$

$$n_{i,t+1} = n_{i,t} + y_{i,t} - y_{i+1,t} \tag{2.19}$$

See Daganzo [Daganzo, 1994] for an alternative derivation of CTM from LWR theory.

With CTM formulated using equation (2.9), the shock waves are subject to spreading when $\frac{w}{v} < 1$, (typical of real world values). To correct this, Daganzo [Daganzo, 1994] offers (2.20) as an improvement to correct the spreading for some types of shocks. See Fig. 2.9 for a comparison showing how this improves shock-wave propagation.

$$y_{i,t} = \min\left\{Q, \quad n_{i-1,t}, \quad \alpha\left(N - n_{i,t}\right)\right\} \tag{2.20}$$

$$\text{where} \quad \alpha = \begin{cases} 1, & \text{if } n_{i-1,t} \le Q \\ \frac{w}{v}, & \text{if } n_{i-1,t} > Q \end{cases}$$

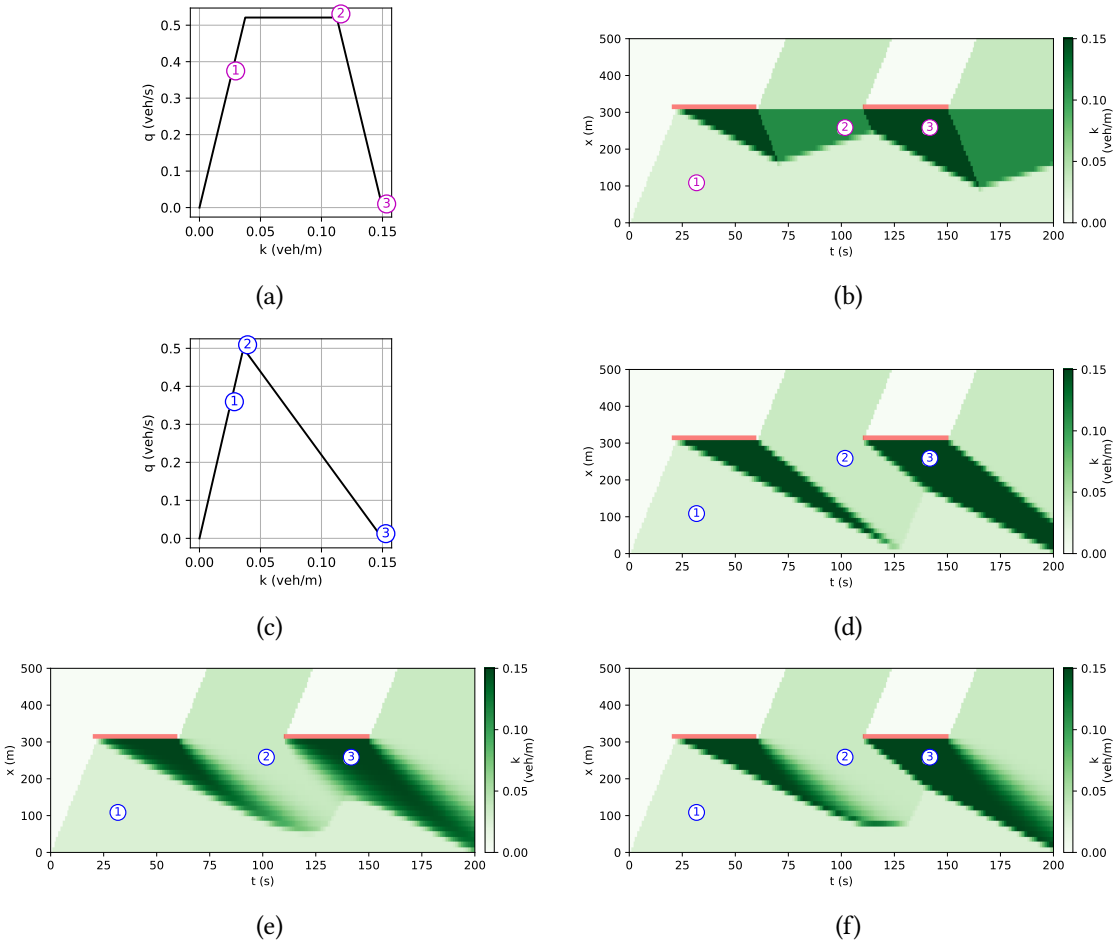


Figure 2.9: Eaxmples of CTM shockwave prediction and spreading. A 500 m section of roadway with a traffic signal at $x = 300$ m, modeled using 36 cells. (a) A trapezoidal fundamental diagram with $\frac{w}{v} = 1$. (b) The cell density over time predicted by CTM using the trapezoidal fundamental diagram, which matches the LWR prediction. (c) A triangular fundamental diagram with $\frac{w}{v} < 1$. (d) The correct cell density predicted by LWR theory for the triangular fundamental diagram. (e) The CTM prediction for the triangular fundamental, exhibiting shockwave spreading. (f) CTM prediction with shockwave spreading partially corrected using (2.20) .

2.5 Newell's Simplified Kinematic Wave Theory

Newell [Newell, 1993a,b,c] derived LWR theory in terms of cumulative flows, and described a method for tracking the shock waves along a section of homogeneous roadway, or link, in terms of the boundary conditions at either end. If it is assumed that $Q()$ in (2.8) is approximated with a triangular fundamental diagram, then there exists at most a single shock wave separating the link into two regimes of flow: congested and free flow (see Fig. 2.7). If points x_0 and x_1 represent the upstream and downstream boundaries of the link, then the boundary conditions of (2.8) at these points are the cumulative arrival and departure curves given by $N(x_0,t) = \int_0^t q(x_0,\tau)d\tau$ and $N(x_1,t) = \int_0^t q(x_1,\tau)d\tau$. Newell then showed that the cumulative function $N(x,t)$ at point (x,t) along a roadway is given by (2.21), where v and w are the forward and backward speed of the shock wave, and ρ^{jam} is the jam density of the link.

$$N(x,t) = \min\left\{N\left(x_0, t - \frac{x - x_0}{v}\right), \quad N\left(x_1, t - \frac{x_1 - x}{w}\right) + (x_1 - x)\rho^{\text{jam}}\right\} \tag{2.21}$$

2.6 The Link Transmission Model (LTM)

Yperman presented the Link Transmission Model (LTM) [Yperman et al., 2005], a link based kinematic wave model using the cumulative arrival and departure curves and applying Newell's theory. By evaluating (2.21) at the beginning and end of the link, over a discrete time interval of Δt, the sending and receiving flows, $S(t)$ and $R(t)$, can be determined at time t with (2.22) and (2.23), where q^c is the maximum flow rate of the link.

$$S(t) = \min\left\{q^c \Delta t, \quad N\left(x_0, t + \Delta t - \frac{x_1 - x_0}{v}\right) - N(x_1, t)\right\} \tag{2.22}$$

$$R(t) = \min\left\{q^c \Delta t, \quad N\left(x_1, t + \Delta t - \frac{x_1 - x_0}{w}\right) + (x_1 - x_0)\rho^{\text{jam}} - N(x_0, t)\right\} \tag{2.23}$$

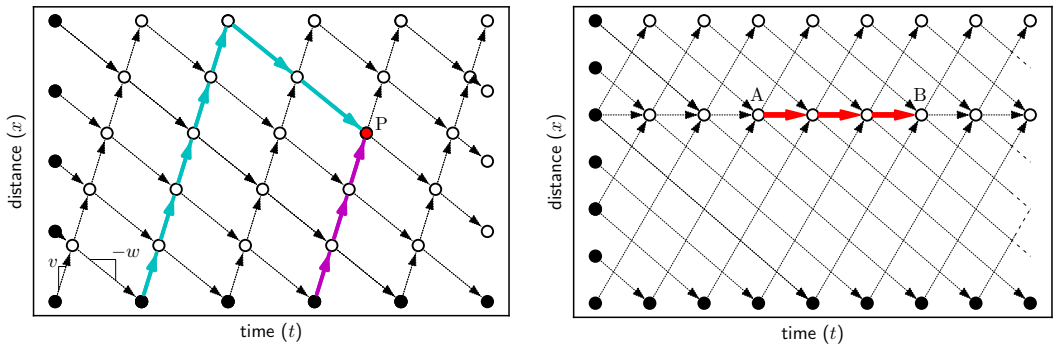
When the shock wave has reached back to the input of the link (spill back), the input of the link is in the congested state and the inflow is determined by the outflow at time $t - \frac{x_1 - x_0}{w}$, the time the backward wave takes to travel the length of the link. Oppositely, if the shock wave has reached forward to the output of the link, then the output is in the free flow state and the outflow is determined by the inflow at time $t - \frac{x_1 - x_0}{v}$, the time the forward wave takes to travel the length of the link. Otherwise the flow is at the maximum rate.

At each time step, the model is used to find the sending and receiving flows based on the current cumulative counts at each end of the link. Once the sending and receiving flows have been determined, the actual number of vehicles transmitted to the downstream links is determined with a merge model, as described in [Yperman et al., 2005].

2.7 Variational Theory (VT)

Daganzo's Variational Theory (VT) [Daganzo, 2005a,b] extends the solution of Newell for solving LWR traffic flows along homogeneous roadways to non-homogeneous roadways with stationary and moving bottlenecks.

Given a set of known boundary conditions **B** along the lines $x = 0$ and $t = 0$, and where:



(a) (b)

Figure 2.10: (a) An example of a *geometric* VT network for a triangular fundamental diagram, showing two of many possible paths from a boundary node (solid dots) to node P. The slope of the edges corresponds to the wave speeds of the fundamental diagram and represents the track of the observer traveling at that speed. (b) A *lopsided* network, where nodes internal to the links have been trimmed for efficiency, and the addition of a *shortcut* path representing a red light signal between nodes A and B.

- $N(t,x)$, is the cumulative vehicle count at location (t,x)
- $\mathbf{P}_\mathrm{P}$ is the set of all paths from the boundary nodes $\mathrm{B} \in \mathbf{B}$ to P
- $\mathbf{V}$ is the set of all valid paths in the network: piece-wise differentiable curves with slopes corresponding to wave speeds in the fundamental diagram.
- $Q(k,t,x)$ is a concave function relating density k to flow q at location (t,x).

Then the solution at location P is given by (2.24) and (2.25),

$$N_\mathrm{P} = \min\{N_\mathrm{B} + \Delta(\mathcal{P})\} : \forall \mathcal{P} \in \mathbf{V} \cap \mathbf{P}_\mathrm{P} \tag{2.24}$$

$$\Delta(\mathcal{P}) = \int_{t_\mathrm{B}}^{t_\mathrm{P}} R(\frac{\mathrm{d}x(t)}{\mathrm{d}t}, t, x)\mathrm{d}t, \quad \text{where } x(t) \text{ is the trajectory of } \mathcal{P} \tag{2.25}$$

And the cost function $R(u,t,x)$ is given by (2.26) and (2.27).

$$u = \frac{\partial}{\partial k} Q(k,t,x) \tag{2.26}$$

$$R(u,t,x) = \sup_k \{Q(k,t,x) - ku\} \tag{2.27}$$

For the special case of a triangular fundamental diagram, u is restricted to the values v, the forward wave speed, and $-w$, the backwards wave speed. Fig. 2.11 shows the relationship between a triangular fundamental diagram for $Q(k,x,t)$ and its cost function $R(u,x,t)$. The problem of finding $N(x,t)$ becomes a shortest path problem on a directed acyclic graph $\mathcal{DG}$ embedded in the (t,x) plane (see Fig. 2.10). The edges of $\mathcal{DG}$ have a slope of v or $-w$ and the nodes are located at points (t,x), separated by Δt and have the value of $N(x,t)$. The solution

for (2.24) then takes the form (2.28),

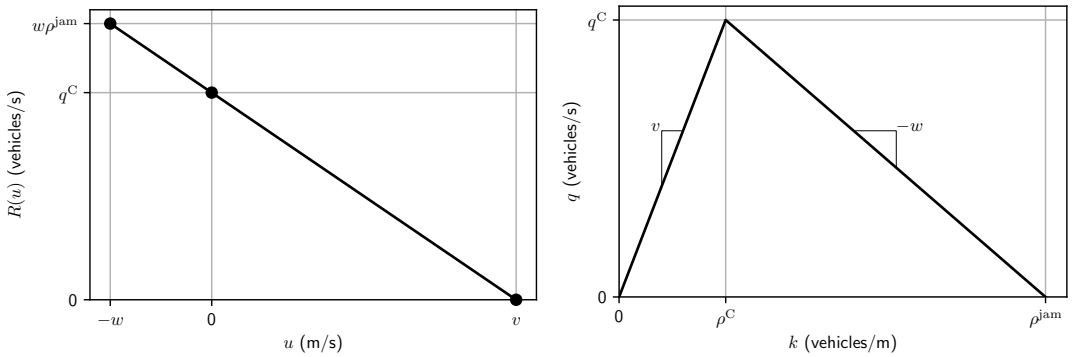
$$N_{\mathrm{P}} = \min\{N_{\mathrm{B}} + \sum_{(i,j)\in\mathcal{P}} c_{ij}\} : \forall \mathcal{P} \in \mathcal{DG} \text{ from B to P}, \forall \mathrm{B} \in \mathbf{B} \tag{2.28}$$

$$c_{ij} = \left(t_j - t_i\right) R(u,t,x) \tag{2.29}$$

$$u = \frac{x_j - x_i}{t_j - t_i} \in \{v, -w\} \tag{2.30}$$

$$R(u,t,x) = \begin{cases} 0, & \text{when } u = v \\ w\rho^{\mathrm{jam}}, & \text{when } u = -w \end{cases} \tag{2.31}$$

where c_{ij}, given by (2.29), is the cost from node i to node j, found by integrating $R(u,t,x)$



(a) (b)

Figure 2.11: (a) The edge cost function for the triangular fundamental diagram in (b), which gives the maximum rate at which vehicles will pass an observer travelling at speed u.

over time along the edge between them, and u, given by (2.30), is the slope of the edge from i to j. Additionally, we can add horizontal edges between nodes with x corresponding to the location of a traffic signal, since the relation $Q(k,t,x) = 0$ for t when the signal is red and no traffic can flow. These are referred to as *shortcuts* in the theory.

The cost function, $R(u,t,x)$, gives the "relative capacity" in vehicles along a wave path and has a physical interpretation is the maximum number of vehicles an observer travelling at the wave speed could pass. With this insight, we see that a cost of zero for a red light enforces that the cumulative count of vehicles passing the stop-line, does not increase from one time step to during the red signal, that is, the relative capacity of a red signal wave path is 0. Similarly, for a triangular fundamental diagram, the cost of the wave path at the free flow speed w is also zero, as the cumulative counts at each end must be the same (no other wave travels faster than v), while for the backwards wave path at the backwards wave speed w, the relative capacity is the number of vehicles that fit in the link at the jam density, $w\rho^{\mathrm{jam}}$, since the downstream cumulative count must not exceed this difference with the upstream count once congestion has spilled back to block input at the upstream end of the link.

2.8 Traffic Signal Optimization

2.8.1 CTM as a MILP

Lo [Lo, 1998] was the first to cast CTM at a MILP and solve for an optimized signal plan. In Lo's formulation, the planning horizon is divided up into T discrete, homogeneous time intervals, and the state of each cell at every time $t \in \{0, 1, 2, 3, \ldots, T\}$ is represented by two variables:

- $n_{i,t} \in [0, N]$, the number of vehicles occupying cell i at time t.
- $y_{i,t} \in [0, Q]$, the flow from cell $i-1$ to i during the interval $[t, t+1)$.

Then, as per Deganzo's CTM formulation, the update for the occupancy of cell i at the end of interval $[t, t+1)$, (given by (2.10)), is constructed directly with constraint (C1).

$$n_{i,t+1} = n_{i,t} + y_{i,t} - y_{i+1,t} \tag{C1}$$

To find the flow $y_{i,t}$, the min $\{Q, \quad n_{i-1,t}, \quad \frac{w}{v}(N - n_{i,t})\}$ of equation (2.9) needs to be evaluated. The upper bound on $y_{i,t}$, is given directly by the upper envelope using constraints (C2) to (C4).

$$y_{i,t} \leq Q \tag{C2}$$
$$y_{i,t} \leq n_{i-1,t} \tag{C3}$$
$$y_{i,t} \leq \frac{w}{v}(N - n_{i,t}) \tag{C4}$$

However, $y_{i,t}$ is still free to take any value between 0 and the upper bound, a condition referred as *traffic withholding* since vehicles may not proceed to the next cell even when there is space for them to do so. To constrain $y_{i,t}$ to the upper bound and prevent withholding, Lo [Lo, 1998] used two additional binary variables, $z^1_{i,t} \in \{0, 1\}$ and $z^2_{i,t} \in \{0, 1\}$ for every cell and time interval, and constraints (C5) to (C7), where $\mathcal{M}$ is a sufficiently large positive constant.

$$Q - \mathcal{M}z^2_{i,t} \leq y_{i,t} \tag{C5}$$
$$n_{i-1,t} - \mathcal{M}z^1_{i,t} - \mathcal{M}(1 - z^2_{i,t}) \leq y_{i,t} \tag{C6}$$
$$\frac{w}{v}(N - n_{i,t}) - \mathcal{M}(1 - z^1_{i,t}) - \mathcal{M}(1 - z^2_{i,t}) \leq y_{i,t} \tag{C7}$$

To model the traffic demand, or inflow to the network, Lo modifies the constraints on the first two cells on every source link in the network. The first cell in a source link has the upper bound constraint on $n_{i,t}$ relaxed, and at $t = 0$ the cell is filled with the the total number of vehicles to enter the network during the planning horizon, that is $n_{i,0} = N^{\text{MAX}}_i$, where $n_{i,t} \in [0, \infty]$ and N^{MAX}_i is the maximum number of cars expected to enter during the time T. The constraints on the flow into the second cell, $y_{i,t}$, are modified to account for the demand profile, by replacing constraints (C2) and (C5), with constraints (C8) and (C9), and $I_{i,t} \in [0, Q]$ is the demand or number of vehicles flowing to the network via cell i at time t.

$$y_{i,t} \leq I_{i,t} \tag{C8}$$
$$I_{i,t} - \mathcal{M}z^2_{i,t} \leq y_{i,t} \tag{C9}$$

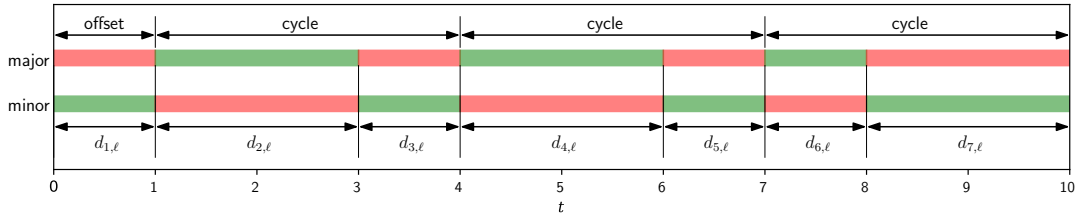


Figure 2.12: A signal plan over 10 time intervals for an intersection with major and minor approaches. The plan has 7 intervals separating 6 signal changes, made up of an initial offset (part cycle), followed by 3 full cycles of the light.

On other words, the first cell acts as a reservoir of vehicles to enter the network, and the flow rate into the second cell meters the vehicles into the network as per the demand profile. This approach allows for elasticity in demand should congestion in the network spill back to the input. Cells at exit points from the network are also modified by having the capacity constraint relaxed, $n_{i,t} \in [0, \infty]$, such that vehicles arriving at the exit cell can flow out unimpeded.

Traffic signals can be modeled by modulating the maximum flow capacity out of the cell preceding the traffic light. A binary variable $p_{\ell,t} \in \{0, 1\}$ is introduced to represent the state of the signal of traffic light ℓ during time interval $[t, t + 1)$, with $p_{\ell,t} = 1$ when the signal is green and 0 when it is red. Additionally, by limiting intersections to simple crossroads, a single binary variable can simultaneously control the side street by inverting the signal. If the flow $y_{i,t}$ from cell $i - 1$ to i is the main approach to light ℓ, and $y_{j,t}$ from cell $j - 1$ to j is the side approach to light ℓ, then constraints (C2) and (C5) are replaced with constraints (C10) and (C11) for the main approach, and constraints (C12) and (C13) for the side approach, for each time interval t.

$$y_{i,t} \leq Q p_{\ell,t} \tag{C10}$$

$$Q p_{\ell,t} - \mathcal{M} z^2_{i,t} \leq y_{i,t} \tag{C11}$$

$$y_{j,t} \leq Q(1 - p_{\ell,t}) \tag{C12}$$

$$Q(1 - p_{\ell,t}) - \mathcal{M} z^2_{j,t} \leq y_{j,t} \tag{C13}$$

Whenever $p_{\ell,t} = 1$, $y_{i,t}$ is free to take on the value determined by (2.9), but when $p_{\ell,t} = 0$ the flow is forced to 0, and inversely for $y_{j,t}$.

A signal plan for a light ℓ represents a series of K signal phases and is described by a set of durations, $\{d_{1,\ell}, \ldots, d_{K,\ell}\}$ (See Fig. 2.12). The initial phase offset is given by $d_{1,\ell} \in [0, \Phi^{\max}_{\ell}]$, and $d_{k,\ell} \in [\Phi^{\min}_{\ell}, \Phi^{\max}_{\ell}]$, for $k >= 2$, represents the duration of phase k. where $\Phi^{\min}$ and $\Phi^{\max}$ are the minimum and maximum bounds for phase duration. Additionally, minimum and maximum cycle times $(\Psi^{\min}_{\ell}, \Psi^{\max}_{\ell})$ can be applied with constraints (C14) and (C15), for $k \in \{2, 4, 6, \ldots\}$.

$$\Psi^{\min}_{\ell} \leq d_{k,\ell} + d_{k+1,\ell} \tag{C14}$$

$$\Psi^{\max}_{\ell} \geq d_{k,\ell} + d_{k+1,\ell} \tag{C15}$$

An additional binary variable, $u_{k,\ell,t} \in \{0,1\}$, is introduced for each phase k to indicate if time t is within the interval $[0, \sum_{j=1}^{k} d_{j,\ell}]$, that is $u_{k,\ell,t} = 1$, if $t \leq \sum_{j=1}^{k} d_{j,\ell}$. The condition for $u_{k,\ell,t}$ can be encoded with constraints (C16) and (C17), where ε is a sufficiently small positive constant.

$$-\mathcal{M}u_{k,\ell,t} + \varepsilon \leq t - d_{k,\ell} \tag{C16}$$

$$t - d_{k,\ell} \leq \mathcal{M}(1 - u_{k,\ell,t}) \tag{C17}$$

Once the time intervals associated with each signal phase have been identified with $u_{k,\ell,t}$, it remains to map the signal variables, $p_{\ell,t}$ in the correct sequence. Lo omits the details, but binary logic for the mapping can be encoded using suitable constraints. For example, to activate the major approach to light ℓ with a green signal during phase 1, a red signal during phase 2, and so on, we could use constraint (C18).

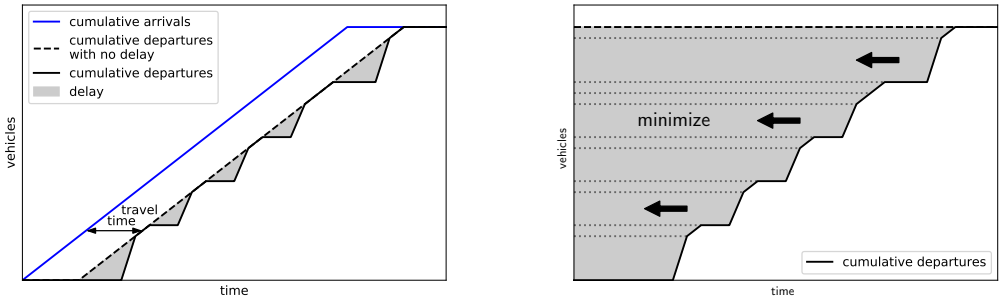
$$p_{\ell,t} = \sum_{k=2}^{K/2} u_{2k,\ell,t} - u_{2k-1,\ell,t} \tag{C18}$$

By making $d_{k,\ell}$ a variable, it is possible for the solver to find an *optimal* signal plan subject to some objective function. Lo proposes minimising total network delay as an objective and determines that after each time step, any vehicle remaining in a cell (i.e. that did not traverse the cell at the free flow speed) experiences a delay of one time step. If $\delta_{i,t} = n_{i,t} - y_{i+1,t}$, then $\delta_{i,t}$ is the number of vehicles remaining in cell i at the end of interval $[t, t+1)$, and represents a measure of delay in vehicle time steps. Then, the total network delay over the planing horizon is the sum of $\delta_{i,t}$ over all cells and time intervals, and the objective is given by constraint (O1).

$$\min \sum_{t} \sum_{i} n_{i,t} - y_{i+1,t} \tag{O1}$$

With an objective of (O1), and given a demand profile, the solver will find an optimal signal plan, subject to the minimum and maximum phase and cycle constraints, that produces the the minimum total network delay under a CTM modelled traffic flow.

Lin and Wang [Lin and Wang, 2004] improve upon Lo's CTM MILP formulation by reducing the total number of binary variables. Rather than solve exactly the min $\{Q,\ n_{i-1,t},\ \frac{w}{v}(N - n_{i,t})\}$ of equation (2.9), they replace constraints (C5) to (C7) with a weighted objective term (2.33) to address the vehicle withholding issue, which removes binary variables $z^1_{i,t}$ and $z^2_{i,t}$ from the formulation. With a suitably small value of β so as not to interfere with the main objective function, the term (2.33), provides an incentive at every time step for vehicles to flow to the next cell, with the upper bound still governed by constraints (C2) to (C4). Lin and Wang dispense with the explicit signal plan of Lo, and instead apply the minimum and maximum constraints for phase and cycle time directly to the signal variable $p_{\ell,t}$, removing the need for the binary variables $y_{k,\ell}$. The moving window summation on the left hand side of equation



(a) (b)

Figure 2.13: Cumulative arrival/departure curves for a network. In (a), the travel time of a vehicle is the horizontal difference between the arrival and departure curves, and delay is any additional time beyond the free flow travel time. Vehicle A encounters delay, while vehicle B does not. The area between the arrival and departure curves is the total travel time for all vehicles in the network. In (b), if the area above the departure curve is minimized, the curve is pulled to the left, and delay is minimized.

(2.32) counts the number of transitions between red and green in the interval $[t, t + \Phi_\ell^{\min}]$.

$$\sum_{\tau=t}^{t+\Phi_\ell^{\min}} |p_{\ell,\tau} - p_{\ell,\tau-1}| \leq 1 \tag{2.32}$$

If the minimum phase constraint $\Phi^{\min}$ is respected, there should be at most one transition during the interval $[t, t + \Phi_\ell^{\min}]$. By introducing the dummy variable, $v_{\ell,t} = |p_{\ell,\tau} - p_{\ell,\tau-1}|$, to indicate a transition, (2.32) can be formulated as constraints (C19) to (C21).

$$p_{\ell,t} - p_{\ell,t-1} \leq v_{\ell,t} \tag{C19}$$

$$p_{\ell,t-1} - p_{\ell,t} \leq v_{\ell,t} \tag{C20}$$

$$\sum_{\tau=t}^{t+\Phi_\ell^{\min}} v_{\ell,\tau} \leq 1 \tag{C21}$$

$$\sum_{\tau=t}^{t+\Phi_\ell^{\max}} p_{\ell,\tau} \leq \Phi_\ell^{\max} \tag{C22}$$

$$p_{\ell,t} - p_{\ell,t+\Psi_\ell^{\text{fixed}}} = 0 \tag{C23}$$

Further, to maintain the maximum phase time, $\Phi_\ell^{\max}$, a similar moving window of duration $[t, t + \Phi_\ell^{\max}]$ can be used and is formulated as constraint (C22). Lin and Wang do not provide constraints for minimum and maximum cycle times, but do provide constraint (C23) to set a fixed cycle time of Ψ_ℓ^{fixed}. The constraint (C23) ensures that the signal pattern repeats with period Ψ_ℓ^{fixed}, and allows for an optimal offset to be selected at the start, but also forces a fixed phase split throughout the planning horizon.

Lin and Wang choose to fix the demand profile, and for cells at the input, constraint (C1)

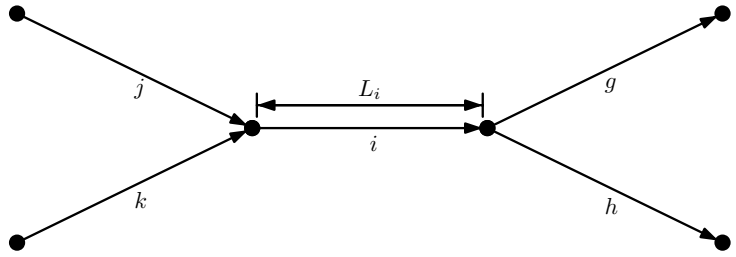


Figure 2.14: An example of an LTM network, with link i having two incoming links, j and k, and two outgoing links, g and h.

is replaced with constraint (C24), and all cell capacities constraints are kept as $n_{i,t} \in [0, N]$.

$$n_{i,t+1} = n_{i,t} + I_{i,t} - y_{i+1,t} \tag{C24}$$

For the outflow from the network an additional variable, $y^{\text{out}}_{i,t}$ is introduced. For cells at the output of the network, at each time step all traffic in the cell is removed via constraint (C25), and constraint (C1) is replaced with constraint (C26).

$$y^{\text{out}}_{i,t} = n_{i,t} \tag{C25}$$

$$n_{i,t+1} = n_{i,t} + y_{i,t} - y^{\text{out}}_{i,t} \tag{C26}$$

Lin and Wang formulate a different objective function than Lo to minimize delay. Their approach is to minimize the total travel time in the network, which is the area between the cumulative arrival and departure curves (see Fig. 2.13). The products $ty^{\text{out}}_{i,t}$ are the areas of the rectangles indicated above the departure curve in Fig. 2.13(b), and their sum forms the area above the cumulative departure curve. Minimizing this area pulls the departure curve to the right and delay in the network is minimized. The full objective function, including the term (2.33) for preventing traffic with-holding, becomes (O2).

$$\beta \sum_i \sum_t t y_{i,t} \tag{2.33}$$

$$\min \sum_i \sum_t t y^{\text{out}}_{i,t} + \beta \sum_i \sum_t t y_{i,t} \tag{O2}$$

2.8.2 LTM as a MILP

The link transmission model was formulated as a MILP by Hajiahmadi et al. [2012]. At each time step t the maximum possible sending and receiving flows between links is determined and the cumulative counts at the entry and exit of each link are updated. Several exogenous parameters are associated with the physical properties of the link and its fundamental diagram:

- $c_i = L_i \rho^{\text{jam}}$, where ρ^{jam} is the jam density of link i in vehicles per meter and L_i is its the length in meters.

- $\mathrm{F}_{i,j}$ is the maximum flow rate of link i and j in vehicles per time step.
- v_i is the integer number of time intervals a forward wave takes to propagate the length of link i, such that $v_i - 1 < \frac{L_i}{v\Delta t} \leq v_i$, where v is the forward wave speed.
- w_i is the integer number of time intervals a backwards wave takes to propagate the length of link i, such that $w_i - 1 < \frac{L_i}{w\Delta t} \leq w_i$, where w is the backwards wave speed.

The formulation uses the following variables:

- $N^{\text{in}}_{i,n}$, the cumulative number of vehicles that have entered link i by the end of interval n,
- $N^{\text{out}}_{i,n}$, the cumulative number of vehicles that have exited link i by the end of interval n,
- $R_{i,n}$, the maximum number of vehicles that can be received by link i during the interval n,
- $S_{i,n}$, the maximum number of vehicles that can be sent by link i during the interval n,
- $\zeta_{i,n}$, the transition number vehicles leaving link i during interval n.

The maximum sending and receiving flows for link i are found by solving the minimums of (2.22) and (2.23) which can be formulated as constraints (L1) to (L8), using the binary variables $z^{\text{in}}_{i,n}$, $z^{\text{out}}_{i,n}$. Further, the sending flow from the link can be modulated by a traffic signal using the binary phase variable $p_{i,j,n}$ which, during interval n, is 1 for a green signal and 0 for a red.

$$R_{i,n} \geq N^{\text{out}}_{i,n-w_i} + c_i - N^{\text{in}}_{i,n-1} - \mathcal{M} z^{\text{in}}_{i,n} \tag{L1}$$
$$R_{i,n} \leq N^{\text{out}}_{i,n-w_i} + c_i - N^{\text{in}}_{i,n-1} \tag{L2}$$
$$R_{i,n} \geq \mathrm{F}_{i,j} - \mathcal{M}(1 - z^{\text{in}}_{i,n}) \tag{L3}$$
$$R_{i,n} \leq \mathrm{F}_{i,j} \tag{L4}$$
$$S_{i,n} \geq N^{\text{in}}_{i,n-v_i} - N^{\text{out}}_{i,n-1} - \mathcal{M} z^{\text{out}}_{i,n} \tag{L5}$$
$$S_{i,n} \leq N^{\text{in}}_{i,n-v_i} - N^{\text{out}}_{i,n-1} \tag{L6}$$
$$S_{i,n} \geq p_{i,j,n}\mathrm{F}_{i,j} - \mathcal{M}(1 - z^{\text{out}}_{i,n}) \tag{L7}$$
$$S_{i,n} \leq p_{i,j,n}\mathrm{F}_{i,j} \tag{L8}$$

For junctions with no turning movements, such that all of the output from link i flows into link j, the transition flow $\zeta_{i,n}$ is given as the minimum of i's sending flow and j's receiving flow. This can be formulated as constraints (L9) to (L12) using the binary variable $\xi_{n,i}$.

$$\zeta_{i,n} \geq S_{i,n} - \mathcal{M}\xi_{n,i} \tag{L9}$$
$$\zeta_{i,n} \leq S_{i,n} \tag{L10}$$
$$\zeta_{i,n} \geq R_{j,n} - \mathcal{M}(1 - \xi_{n,i}) \tag{L11}$$
$$\zeta_{i,n} \leq R_{j,n} \tag{L12}$$

Finally the transition update functions for the cumulative counts at the link boundaries can be solved by (L13) and (L14).

$$N^{\text{out}}_{i,n} = N^{\text{out}}_{i,n-1} + \zeta_{i,n} \tag{L13}$$
$$N^{\text{in}}_{i,n} = N^{\text{in}}_{i,n-1} + \zeta_{j,n} \tag{L14}$$

For junctions with turning movements, the transition flow sent from one link must be distributed among the receiving links. While Hajiahmadi et al. [2012] omits details, we can use a simple model using turn fractions such that $\text{Pr}_{i,j}$ gives the proportion of vehicles leaving link i that enter link j, and the sum of all turn fractions from a link is 1. For the case of one sending link, i, into two receiving links, j and k, the transition flow is the minimization $\zeta_{i,n} = \min\left\{S_{i,n}, \frac{R_{j,n}}{\text{Pr}_{i,j}}, \frac{R_{k,n}}{\text{Pr}_{i,k}}\right\}$, and this is formulated as constraints (L15) to (L20).

$$\zeta_{i,n} \geq S_{i,n} - \mathcal{M}\eta_{n,i} - \mathcal{M}\xi_{n,i} \tag{L15}$$
$$\zeta_{i,n} \leq S_{i,n} \tag{L16}$$
$$\zeta_{i,n} \geq R_{j,n}\frac{1}{\text{Pr}_{i,j}} - \mathcal{M}\eta_{n,i} - \mathcal{M}(1-\xi_{n,i}) \tag{L17}$$
$$\zeta_{i,n} \leq R_{j,n}\frac{1}{\text{Pr}_{i,j}} \tag{L18}$$
$$\zeta_{i,n} \geq R_{k,n}\frac{1}{\text{Pr}_{i,k}} - \mathcal{M}(1-\eta_{n,i}) - \mathcal{M}(1-\xi_{n,i}) \tag{L19}$$
$$\zeta_{i,n} \leq R_{k,n}\frac{1}{\text{Pr}_{i,k}} \tag{L20}$$

Such a model does not support the simultaneous merging of traffic flows, but this is acceptable if we consider that all vehicle movements through an intersection are protected by a traffic signal.

2.8.3 Variational Theory as a MILP

Wada [Wada et al., 2017] was the first to cast Variational Theory as a MILP, and showed how to find an optimal signal policy by solving a discrete shortest path problem on a lopsided VT network. Wada formulates the objective of minimizing the total delay in the network with (2.34). The delay can be seen as the difference between the total cumulative departure curve and the total cumulative departure curve with no delay. The total cumulative departure curve with no delay is the total cumulative departures shifted by the free flow travel time.

$$D(\mathbf{s}) = \sum_t N(t - x_{\text{exit}}/u, 0)\Delta t - \sum_t N(t, x_{\text{exit}})\Delta t \tag{2.34}$$
$$D(\mathbf{s}) = U - \sum_{j \in \mathcal{V}_{\text{exit}}} N_j \Delta t \tag{2.35}$$

$$\min D(\mathbf{s}) = \min U - \sum_{j \in \mathcal{V}_{\text{exit}}} N_j \Delta t = U - \max \sum_{j \in \mathcal{V}_{\text{exit}}} N_j \Delta t \tag{2.36}$$

$$\implies \max \sum_{j \in \mathcal{V}_{\text{exit}}} N_j \Delta t \tag{2.37}$$

The network consists of nodes and links in space–time, with nodes denoted by the set $\mathcal{V}$. Links are divided into two sets: ordinary links with a slope of u or $-w$ dentoted $\mathcal{L}_{\text{o}}$, and signal links with slope of 0 denoted $\mathcal{L}_{\text{s}}$. Nodes are identified by an index $i = (t_i, x_i)$ and links are identified by their nodes as a pair (i, j), where (i, j) is the link directed from i to j. Nodes at the network exit are denoted by the set $\mathcal{V}_{\text{exit}}$. The cost of link (i, j) is represented by c_{ij} and is given by Variatonal Theory (refer to Fig. 2.11(a)). For ordinary links representing forward waves with slope u, $c_{ij} = 0$, while backwave links with slope $-w$, $c_{ij} = (t_j - t_i) w \rho^{\text{jam}}$. Traffic signals are represented with horizontal shortcut links with $c_{ij} = p_{ij}(t_j - t_i) q^{\text{C}}$, where $p_{ij} \in \{0, 1\}$ is the signal state between nodes i and j. If the signal is green then $c_{ij} = (t_j - t_i) q^{\text{C}}$ and when the signal is red no traffic can flow and $c_{ij} = 0$ since $Q(k, t, x) = 0$ for $t_i \leq t < t_j$. To cast the problem as a single origin, multiple destination shortest path problem, a dummy origin node o is added to $\mathcal{V}$ and links from o to each node at the boundary, with cost $c_{oi} = N_i$, where N_i is the known cumulative count at boundary node i. To find the value of N_i for $i \in \mathcal{V}_{\text{exit}}$ the problem can now be solved using the linear program formulation of the shortest path problem given by constraints (O3) and (V1).

$$U - \min_{\mathbf{y} \geq \mathbf{0}} \sum_{(i,j) \in \mathcal{L}_{\text{o}} \cup \mathcal{L}_{\text{s}}} c_{ij} y_{ij} \tag{O3}$$

$$\text{s.t.} \quad \sum_j y_{ji} - \sum_j y_{ij} = \delta_{id} \quad \forall i \in \mathcal{V} \setminus \{o\} \tag{V1}$$

where $\delta_{id} = 1$ if $i = d \in \mathcal{V}_{\text{exit}}$, else 0

Note, however, that the constraint (O3) contains the product of p_{ij} and y_{ij}, and if p_{ij} is also a variable to be solved for the optimal signal policy, then (O3) and (V1) would be a bi-linear problem. To avoid this Wada proposed using the dual form of the shortest path program given by constraints (O4) and (V2).

$$\max_N \sum_{j \in \mathcal{V}_{\text{exit}}} N_j \Delta t \tag{O4}$$

$$\text{s.t.} \quad N_j \Delta t \leq N_i \Delta t + c_{ij} \quad \forall (i, j) \in \mathcal{L}_{\text{o}} \cup \mathcal{L}_{\text{s}} \tag{V2}$$

Signal Constraint Network

Additionally Wada [Wada et al., 2017] defines a Signal Constraint Network (SCN) to enforce signal timing constraints and phase sequencing. An SCN for intersection ℓ is a set of nodes, $\mathcal{V}^\ell$, and links, $\mathcal{L}^\ell$ as shown in Fig. 2.15. Further, $\mathcal{L}^\ell$ is made up of *signal links* that represent the state of a traffic signal over the time interval spanned by the link, and *dummy links* used to enforce the timing constraints for minimum green time, lost time and optionally a predefined fixed cycle time. All links in the SCN have zero cost and an origin node and destination node, o and d, are connected to the network with additional dummy links. The signal timing constraints are satisfied when a unique path is found from o to d. To find a valid solution on a

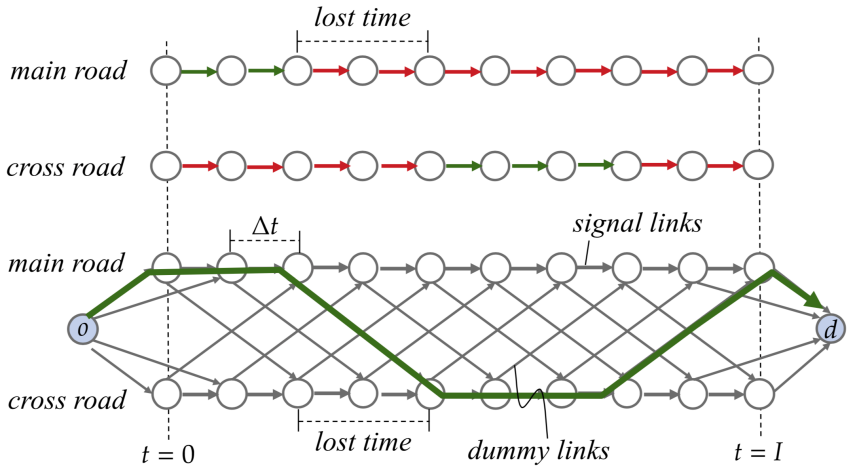


Figure 2.15: A Signal Constraint Network with lost time = $2\Delta t$.
(From Figure 3, Wada et al. [2017]).

SCN we can use a similar method to the primal shortest path problem given in (V1). Each link, (i, j), in the SCN is represented with a binary variable, z_{ij}^{ℓ} and is formulated as constraints with (V3).

$$\sum_j z_{ji}^{\ell} - \sum_j z_{ij}^{\ell} = \delta_{id} \quad \forall i \in \mathcal{V}^{\ell} \setminus \{o\} \tag{V3}$$

where $\delta_{id} = 1$ if $i = d \in \mathcal{V}^{\ell}$, else 0

For all nodes that are not o or d, constraint (V3) enforces either no flow through the node, or exactly one in-flowing link and one out-flowing link, while nodes o and d must have exactly one out-flowing and in-flowing link respectively. There is a linear mapping between the SCN variables z_{ij}^{ℓ} and VT signal variables p_{ij}, given by a transform matrix $\mathbf{T}$. The SCN signal variables are arranged into is a vector $\vec{z}^{\ell}$, which is split into sub-vectors of signal links, $\vec{z}_{\mathrm{s}}^{\ell}$, and dummy links, $\vec{z}_{\mathrm{d}}^{\ell}$. Similarly the signal variables are arranged into a vector $\vec{p}^{\ell}$ split into subvectors for each phase of intersection ℓ. The transform is given by constraint (2.38), where $\mathbf{T} = \begin{bmatrix}\mathbf{I} & \mathbf{B}^{\ell}\end{bmatrix}$, and $\mathbf{B}^{\ell}$ is the mapping between dummy links in $\vec{z}_{\mathrm{d}}^{\ell}$ representing minimum green time and signal variables in $\vec{p}^{\ell}$.

$$\begin{bmatrix}\vec{p}_k^{\ell} \\ \vec{p}_{k+1}^{\ell}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{B}^{\ell}\end{bmatrix} \begin{bmatrix}\vec{z}_{\mathrm{s}}^{\ell} \\ \vec{z}_{\mathrm{d}}^{\ell}\end{bmatrix} \tag{2.38}$$

The transform (2.38) results in a set of constraints given by (V4).

$$p_{ij} = z_{ij}^{\ell} + \sum_{(i',j') \in \mathbf{B}^{\ell}} z_{i'j'}^{\ell} \tag{V4}$$

2.8.4 LKWM as a MILP

Han [Han et al., 2012, 2016a,b] formulated the LKWM as a MILP. Similar to Lo's CTM formulation, the planing horizon T is divided up into N homogeneous intervals each of Δt duration. Han models the road network as a series of links connected at signal controlled intersections. Turning probabilities are introduced to model the weighting of divergent flows out of a link and into at most 2 downstream links. For simplicity the formulation assumes that only one link will flow into a junction at a time, and this is enforced by the signal phase sequencing. Each link i then has the following properties:

- $Q_i = \rho^{\text{jam}} L_i$: the maximum number of vehicles that can fit within link i of length L_i at the jam density ρ^{jam}
- F_i: the maximum capacity flow rate of link i
- $Pr_{i,j}$: the turn probability from link i to j
- $T_i^{\text{p}} = \frac{L_i}{v}$: the time for forward wave to propagate the length of link i at speed v
- $T_i^{\text{w}} = \frac{L_i}{w}$: the time for backwards wave to propagate the length of link i at speed w
- $\Delta_i^{\text{p}} = \left\lfloor \frac{T_i^{\text{p}}}{\Delta t} \right\rceil$: the number of time intervals for the forward wave to propagate link i, rounded to the nearest integer
- $\Delta_i^{\text{w}} = \left\lfloor \frac{T_i^{\text{w}}}{\Delta t} \right\rceil$: the number of time intervals for the backwards wave to propagate link i, rounded to the nearest integer

And associated with each link is a set of variables:

- $q_{i,n}^{\text{in}} \in [0, F_i]$: the flow rate of vehicles into link i during interval n.
- $q_{i,n}^{\text{out}} \in [0, F_i]$: the flow rate of vehicles out of link i during interval n.
- $f_{i,n}^{\text{in}} \in [0, F_i]$: the maximum flow rate of vehicles into link i during interval n.
- $f_{i,n}^{\text{out}} \in [0, F_i]$, the maximum flow rate of vehicles out of link i during interval n.
- $r_{i,n}^{\text{in}} \in \{0, 1\}$: the traffic flow state at the input of link i during interval n.
- $r_{i,n}^{\text{out}} \in \{0, 1\}$: the traffic flow state at the output of link i during interval n.
- $p_{i,n} \in \{0, 1\}$: the state of the traffic signal controlling link i during interval n.

The flow state variables $r_{i,n}^{\text{in}}$ and $r_{i,n}^{\text{our}}$ indicate where the input and output, respectively, are congested:

$$r_{i,n}^{\text{in}} = \begin{cases} 1, & \text{if } \Delta t \sum_{k=1}^{n-\Delta_i^{\text{w}}} q_{i,k}^{\text{out}} + Q_i \leq \Delta t \sum_{k=1}^{n} q_{i,k}^{\text{in}} \\ 0, & \text{otherwise} \end{cases} \tag{2.39}$$

$$r_{i,n}^{\text{out}} = \begin{cases} 0, & \text{if } \Delta t \sum_{k=1}^{n-\Delta_i^{\text{p}}} q_{i,k}^{\text{in}} \leq \Delta t \sum_{k=1}^{n} q_{i,k}^{\text{out}} \\ 1, & \text{otherwise} \end{cases} \tag{2.40}$$

In the equations (2.39) and (2.39), the states of $r_{i,n}^{\text{in}}$ and $r_{i,n}^{\text{out}}$ are determined from the cumulative arrival and departure curves using Newell's theory of link based kinematic waves (2.21), and this is translated to constraints (H1) to (H4).

$$\Delta t \sum_{k=1}^{n-\Delta_i^{\text{w}}} q_{i,k}^{\text{out}} - \Delta t \sum_{k=1}^{n} q_{i,k}^{\text{in}} + Q_i \leq \mathcal{M}(1 - r_{i,n}^{\text{in}}) + \varepsilon \tag{H1}$$

$$\Delta t \sum_{k=1}^{n-\Delta_i^{\text{w}}} q_{i,k}^{\text{out}} - \Delta t \sum_{k=1}^{n} q_{i,k}^{\text{in}} + Q_i \geq -\mathcal{M} r_{i,n}^{\text{in}} + \varepsilon \tag{H2}$$

$$\Delta t \sum_{k=1}^{n-\Delta_i^{\text{p}}} q_{i,k}^{\text{in}} - \Delta t \sum_{k=1}^{n} q_{i,k}^{\text{out}} \leq \mathcal{M} r_{i,n}^{\text{out}} + \varepsilon \tag{H3}$$

$$\Delta t \sum_{k=1}^{n-\Delta_i^{\text{p}}} q_{i,k}^{\text{in}} - \Delta t \sum_{k=1}^{n} q_{i,k}^{\text{out}} \geq \mathcal{M}(r_{i,n}^{\text{out}} - 1) + \varepsilon \tag{H4}$$

Next, the upper bounds on the input and output flow rates, $f_{i,n}^{\text{in}}$ and $f_{i,n}^{\text{out}}$, are determined from flow states, $r_{i,n}^{\text{in}}$ and $r_{i,n}^{\text{out}}$, using equations (2.41) and (2.42).

$$f_{i,n}^{\text{in}} = \begin{cases} F_i, & \text{if } r_{i,n}^{\text{in}} = 0 \\ q_{i,n-\Delta_i^{\text{w}}}^{\text{out}}, & \text{if } r_{i,n}^{\text{in}} = 1 \end{cases} \tag{2.41}$$

$$f_{i,n}^{\text{out}} = \begin{cases} F_i, & \text{if } r_{i,n}^{\text{out}} = 1 \\ q_{i,n-\Delta_i^{\text{p}}}^{\text{in}}, & \text{if } r_{i,n}^{\text{out}} = 0 \end{cases} \tag{2.42}$$

Which can be cast as constraints (H6) to (H8).

$$F_i + \mathcal{M}(r_{i,n}^{\text{out}} - 1) \leq f_{i,n}^{\text{out}} \leq F_i \tag{H5}$$

$$q_{i,n-\Delta_i^{\text{p}}}^{\text{in}} - \mathcal{M} r_{i,n}^{\text{out}} \leq f_{i,n}^{\text{out}} \leq q_{i,n-\Delta_i^{\text{p}}}^{\text{in}} + \mathcal{M} r_{i,n}^{\text{out}} \tag{H6}$$

$$F_i + \mathcal{M} r_{i,n}^{\text{in}} \leq f_{i,n}^{\text{in}} \leq F_i \tag{H7}$$

$$q_{i,n-\Delta_i^{\text{w}}}^{\text{out}} + \mathcal{M}(r_{i,n}^{\text{in}} - 1) \leq f_{i,n}^{\text{in}} \leq q_{i,n-\Delta_i^{\text{w}}}^{\text{out}} - \mathcal{M}(r_{i,n}^{\text{in}} - 1) \tag{H8}$$

The input and output flows $q_{i,n}^{\text{in}}$, $q_{i,n}^{\text{out}}$, during interval n, can now be determined from $f_{i,n}^{\text{in}}$ and $f_{i,n}^{\text{out}}$. Firstly, the output flow $q_{i,n}^{\text{out}}$ is the minimum flow of equation (2.43), weighted by the traffic signal state.

$$q_{i,n}^{\text{out}} = p_{i,n} \min \left\{ f_{i,n}^{\text{out}}, \quad \frac{f_{j,n}^{\text{in}}}{\text{Pr}_{i,j}}, \quad \frac{f_{k,n}^{\text{in}}}{\text{Pr}_{i,k}} \right\} \tag{2.43}$$

By introducing helper variable $\xi_{i,n}$ and binary variables $z^1_{i,n} \in \{0,1\}$ and $z^2_{i,n} \in \{0,1\}$, equation (2.43) can be cast as constraints (H9) to (H13).

$$f^{\text{out}}_{i,n} - \mathcal{M} z^2_{i,t} \leq \xi_{i,n} \leq f^{\text{out}}_{i,n} \tag{H9}$$

$$\frac{f^{\text{in}}_{j,n}}{\text{Pr}_{i,j}} - \mathcal{M} z^1_{i,n} - \mathcal{M}(1 - z^2_{i,n}) \leq \xi_{i,n} \leq \frac{f^{\text{in}}_{j,n}}{\text{Pr}_{i,j}} \tag{H10}$$

$$\frac{f^{\text{in}}_{k,n}}{\text{Pr}_{i,k}} - \mathcal{M}(1 - z^1_{i,n}) - \mathcal{M}(1 - z^2_{i,n}) \leq \xi_{i,n} \leq \frac{f^{\text{in}}_{k,n}}{\text{Pr}_{i,k}} \tag{H11}$$

$$0 \leq q^{\text{out}}_{i,n} \leq \mathcal{M} p_{i,n} \tag{H12}$$

$$\xi_{i,n} + \mathcal{M}(p_{i,n} - 1) \leq q^{\text{out}}_{i,n} \leq \xi_{i,n} \tag{H13}$$

Finally, $q^{\text{in}}_{i,n}$ is the sum of the output flows of the links j and k, flowing into i, weighted by the turn probabilities, $\text{Pr}_{k,i}$ and $\text{Pr}_{k,i}$, and is given in constraint (H14).

$$q^{\text{in}}_{i,n} = \text{Pr}_{j,i} q^{\text{out}}_{j,n} + \text{Pr}_{k,i} q^{\text{out}}_{k,n} \tag{H14}$$

To ensure that only one signal phase is active at a time in any intersection, constraint (H15) is introduced, where i and j are the two approaches to the intersection. Constraint (H15) ensures that the is only one inflow at a time to an intersection, avoiding the need to model merging flows

$$p_{i,n} + p_{j,n} = 1 \tag{H15}$$

Han makes no attempt to constrain the phase or cycle durations for the signals, pointing out that this results in the most flexible phase splits and cycle durations, although this ignores practical considerations such as a minimum green time needed for cross walk signals, and unbounded waits with no upper bound on red signal duration.

As an objective function Han proposed to maximise flow as the inverse of delay, using objective (O5).

$$\max \sum_{n=1}^{N} \frac{1}{n+1} \sum_{i \in \mathcal{I}} q^{\text{out}}_{i,n} \tag{O5}$$

Here, $\mathcal{I}$ is the set of links prioritized in the solution, and the weighting $\frac{1}{n+1}$ is an incentive for solutions to maximize traffic flow as early as possible within the planning horizon.

2.9 Summary

In this chapter we looked at basic traffic flow theory, before deriving the Hamilton Jacobi form of the LWR equation using the fundamental relation of flow and density.

CTM was the first formulation to offer a stable method of solving the LWR equation and simulating traffic flow in roadways with the modeling of shock wave phenomena. Newell observed that the boundary conditions of a homogeneous link can be used to find the function $N(x,t)$, from the constraints relating the cumulative arrival and departures curves. And later

Yperman extend Newell's theory with the LTM formulation to model LWR traffic flows on networks of homogeneous links, using a triangular fundamental diagram. Finally Deganzo showed that varational theory can be used to solve the Hamilton Jacobi form of the LWR equation, by solving a shortest path problem on lattice of wave paths embedded in a space-time diagram.

CTM was first cast as a MILP for traffic signal optimization by Lo along with an objective function for minimizing delay, and was later extended by Lin and Wang. LTM was cast as a MILP by Hajiahmadi for a ramp metering application, but received little further interest. Wada first formulated varational theory as a MILP using the dual form of the shortest path LP, and gave an extension for stochastic arrivals.

There are several useful things to consider with these formulations:

1. Link based approaches use fewer variables for the same number of time steps, than CTM.

2. Lin and Wang's trick of using a withholding term in the objective function to avoid direct evaluation of minimizations, saves the use of additional binary variables.

3. The VT formulation of Wada does not support turning movements

We will make use of these observations in the following chapters.

Chapter 3

The Queue Transmission Model

3.1 Overview

In this chapter, we build on the body of work in mixed integer linear programming (MILP) approaches that attempt to jointly optimize traffic signal control over an *entire traffic network* (rather than focus on arterial routes) and specifically on improving the scalability of these methods for large urban traffic networks. In our investigation of existing approaches in this vein, namely exemplar methods in the spirit of [Lo, 1998; Lin and Wang, 2004] that use a (modified) cell transmission model (CTM) [Daganzo, 1994, 1995] for their underlying prediction of traffic flows, we remark that a major drawback is the CTM-imposed requirement to choose a predetermined *homogeneous* (and often necessarily small) time step for reasonable modeling fidelity. This need to model a large number of CTM cells with a small time step leads to MILPs that are exceedingly large and often intractable to solve.

Our primary insight in this work stems from the fact that MILP-based approaches to traffic control used in a receding horizon control manner (that replan at fixed time intervals) need to compute high fidelity control policies only for the early stages of the signal plan; therefore, coarser time steps can be employed to "see" over a long horizon to preemptively adapt to distant platoons and other predicted long-term changes in traffic flows. This need for non-homogeneous control in turn spawns the need for an additional innovation: we require a traffic flow model that permits non-homogeneous time steps and properly models the travel time delay between lights. To this end, we might consider CTM extensions such as the variable cell length CTM [Xiaojian et al., 2010], stochastic CTM [Sumalee et al., 2011; Jabari and Liu, 2012], CTM extensions for better modeling freeway-urban interactions [Huang, 2011] including CTM hybrids with link-based models [Muralidharan et al., 2009], assymmetric CTMs for better handling flow imbalances in merging roads [Gomes and Horowitz, 2006], the situational CTM for better modeling of boundary conditions [Kim, 2002], and the lagged CTM for improved modeling of the flow density relation [Lu et al., 2011]. However, despite the widespread varieties of the CTM and usage for a range of applications [Alecsandru et al., 2011], there seems to be no extension that permits *non-homogeneous* time steps as proposed in our novel MILP-based control approach.

For this reason, as a major contribution of this work to enable our non-homogeneous time MILP-based model of joint intersection control, we contribute the queue transmission model (QTM) that blends elements of cell-based and link-based modeling approaches as illustrated and summarized in Figure 3.1. The QTM offers the following key benefits:

- Unlike previous CTM-based joint intersection signal optimization [Lo, 1998; Lin and Wang, 2004; Islam et al., 2020], the QTM is intended for *non-homogeneous* time steps that can be used for control over large horizons.

- Any length of roadway without merges or diverges can be modeled as a single queue leading to compact QTM MILP encodings of large traffic networks (i.e., large numbers of cells and their associated MILP variables are not required *between* intersections). Further, the free flow travel time of a link can be modeled exactly, independent of the discritizaiton time step, while CTM requires a further increased discretization to approach the same resolution.

- The QTM accurately models fixed travel time delays critical to green wave coordination

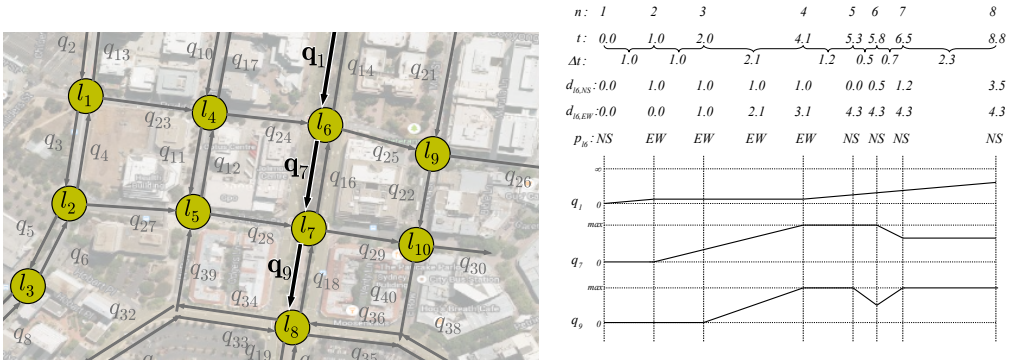


Figure 3.1: (a) Example of a real traffic network modeled using the QTM. (b) A preview of different QTM model parameters as a function of *non-homogeneous* discretized time intervals indexed by n. For each n, we show the following parameters: the elapsed time t, the non-homogeneous time step length Δt, the cumulative duration d of two different light phases for l_6, the phase p of light l_6, and the traffic volume of different queues q linearly interpolated between time points. There is technically a binary p for each phase, but we abuse notation and simply show the current active phase: *NS* for *north-south green* and *EW* for *east-west green* assuming the top of the map is north. Here we see that traffic progresses from q_1 to q_7 to q_9 according to light phases and traffic propagation delay with non-homogeneous time steps only at required changepoints. We refer to the QTM model section for precise notation and technical definitions.

as in [Gartner et al., 1974; Gartner and Stamatiadis, 2002; He et al., 2011] through the use of a non-first order Markovian update model and further combines this with fully joint intersection signal optimization in the spirit of [Lo, 1998; Lin and Wang, 2004; Han et al., 2012].

In the remainder of this chapter, we first formalize our novel QTM model of traffic flow with non-homogeneous time steps and show how to encode it as a linear program for computing traffic flows. Next we proceed to allow the traffic signals to become discrete phase variables that are optimized subject to a delay minimizing objective and standard minimum and maximum time constraints for cycles and phases; this results in our final MILP formulation of traffic signal control. We then experiment with this novel QTM-based MILP control in a range of traffic networks and demonstrate that the non-homogeneous MILP formulation achieves (i) substantially lower delay solutions, (ii) improved per-vehicle delay distributions, and (iii) more optimal travel times over a longer horizon in comparison to the homogeneous MILP formulation with the same number of binary and continuous variables.

3.2 The Queue Transmission Model (QTM)

A Queue Transmission Model (QTM) is the tuple $(\mathcal{Q}, \mathcal{L}, \vec{\Delta t}, \mathbf{I})$, where $\mathcal{Q}$ and $\mathcal{L}$ are, respectively, the set of queues and lights; $\vec{\Delta t}$ is a vector of size N representing the homogeneous, or non-homogeneous, discretization of the problem horizon $[0, \mathrm{T}]$ and the duration in seconds

of the n-th time interval is denoted as Δt_n; and $\mathbf{I}$ is a matrix $|\mathcal{Q}| \times \mathrm{T}$ in which $I_{i,n}$ represents the flow of vehicles requesting to enter queue i from the outside of the network at time n.

A **traffic light** $\ell \in \mathcal{L}$ is defined as the tuple $(\Psi_\ell^{\min}, \Psi_\ell^{\max}, \mathcal{P}_\ell, \vec{\Phi}_\ell^{\min}, \vec{\Phi}_\ell^{\max})$, where:

- $\mathcal{P}_\ell$ is the set of phases of ℓ;
- $\Psi_\ell^{\min}$ ($\Psi_\ell^{\max}$) is the minimum (maximum) allowed cycle time for ℓ; and
- $\vec{\Phi}_\ell^{\min}$ ($\vec{\Phi}_\ell^{\max}$) is a vector of size $|\mathcal{P}_\ell|$ and $\Phi_{\ell,k}^{\min}$ ($\Phi_{\ell,k}^{\max}$) is the minimum (maximum) allowed time for phase $k \in \mathcal{P}_\ell$.

A **queue** $i \in \mathcal{Q}$ represents a segment of road that vehicles traverse at free flow speed; once traversed, the vehicles are vertically stacked in a stop line queue. Formally, a queue i is defined by the tuple $(\mathrm{Q}_i, \mathrm{T}_i^{\mathrm{p}}, \mathrm{F}_i^{\mathrm{out}}, \vec{F}_i, \vec{Pr}_i, \mathcal{Q}_i^{\mathcal{P}})$ where:

- Q_i is the maximum capacity of i;
- $\mathrm{T}_i^{\mathrm{p}}$ is the time required to traverse i and reach the stop line;
- $\mathrm{F}_i^{\mathrm{out}}$ represents the maximum traffic flow from i to the outside of the modeled network;
- $\vec{F}_i$ and $\vec{Pr}_i$ are vectors of size $|\mathcal{Q}|$ and their j-th entry (i.e., $\mathrm{F}_{i,j}$ and $\mathrm{Pr}_{i,j}$) represent the maximum flow from queue i to j and the turn probability from i to j (where $\sum_{j \in \mathcal{Q}} \mathrm{Pr}_{i,j} = 1$), respectively; and
- $\mathcal{Q}_i^{\mathcal{P}}$ is the set of traffic light phases controlling the outflow of queue i, where the pair, $(\ell, k) \in \mathcal{Q}_i^{\mathcal{P}}$, denotes phase k of light ℓ.

Differently than the CTM [Daganzo, 1994; Lin and Wang, 2004], the QTM does not assume that $\Delta t_n = \mathrm{T}_i^{\mathrm{p}}$ for all n, that is, the QTM can represent non-homogeneous time intervals (Section 3.1). The only requirement over Δt_n is that no traffic light maximum phase time is smaller than any Δt_n since phase changes occur only between time intervals; formally, $\Delta t_n \leq \min_{\ell \in \mathcal{L}, k \in \mathcal{P}_\ell} \Phi_{\ell,k}^{\max}$ for all $n \in \{1, \ldots, \mathrm{N}\}$.

3.2.1 Computing Traffic Flows with QTM

In this section, we present how to compute traffic flows using QTM and non-homogeneous time intervals Δt. We assume for the remainder of this section that a *valid* control plan for all traffic lights is fixed and given as parameter; formally, for all $\ell \in \mathcal{L}$, $k \in \mathcal{P}_\ell$, and interval $n \in \{1, \ldots, \mathrm{N}\}$, the binary variable $p_{\ell,k,n}$ is known a priori and indicates if phase k of light ℓ is active (i.e., $p_{\ell,k,n} = 1$) or not on interval n. Each phase $k \in \mathcal{P}_\ell$ can control the flow from more than one queue, allowing arbitrary intersection topologies to be modelled, including "all red" phases as a switching penalty and modeling lost time from amber lights.

We represent the problem of finding the maximal flow between capacity-constrained queues as a Linear Program (LP) over the following variables defined for all intervals $n \in \{1, \ldots, \mathrm{N}\}$ and queues i and j:

- $q_{i,n} \in [0, \mathrm{Q}_i]$: traffic volume waiting in the stop line of queue i at the beginning of interval n;
- $f^{\text{in}}_{i,n} \in [0, I_{i,n}]$: inflow to the network via queue i during interval n;
- $f^{\text{out}}_{i,n} \in [0, \mathrm{F}^{\text{out}}_i]$: outflow from the network via queue i during interval n; and
- $f_{i,j,n} \in [0, \mathrm{F}_{i,j}]$: flow from queue i into queue j during interval n.

The maximum traffic flow from queue i to queue j is enforced by constraints (Q1) and (Q2). (Q1) ensures that only the fraction $\mathrm{Pr}_{i,j}$ of the total internal outflow of i goes to j, and since each $f_{i,j,n}$ appears on both sides of (Q1), the upstream queue i will block if any downstream queue j is full. (Q2) forces the flow from i to j to be zero if all phases controlling i are inactive (i.e., $p_{\ell,k,n} = 0$ for all $(\ell, k) \in \mathcal{Q}^{\mathcal{P}}_i$). If more than one phase $p_{\ell,k,n}$ is active, then (Q2) is subsumed by the domain upper bound of $f_{i,j,n}$.

$$f_{i,j,n} = \mathrm{Pr}_{i,j} \sum_{k=1}^{|\mathcal{Q}|} f_{i,k,n} \tag{Q1}$$

$$f_{i,j,n} \leq \mathrm{F}_{i,j} \sum_{(\ell,k) \in \mathcal{Q}^{\mathcal{P}}_i} p_{\ell,k,n} \tag{Q2}$$

To simplify the presentation of the remainder of the LP, we define the helper variables $q^{\text{in}}_{i,n}$ (Q3), $q^{\text{out}}_{i,n}$ (Q4), and t_n (3.1) to represent the volume of traffic to enter and leave queue i during interval n, and the time elapsed since the beginning of the problem until the end of interval Δt_n, respectively.

$$q^{\text{in}}_{i,n} = \Delta t_n (f^{\text{in}}_{i,n} + \sum_{j=1}^{|\mathcal{Q}|} f_{j,i,n}) \tag{Q3}$$

$$q^{\text{out}}_{i,n} = \Delta t_n (f^{\text{out}}_{i,n} + \sum_{j=1}^{|\mathcal{Q}|} f_{i,j,n}) \tag{Q4}$$

$$t_n = \sum_{x=1}^{n} \Delta t_x \tag{3.1}$$

In order to account for the misalignment of the different Δt and $\mathrm{T}^{\mathrm{p}}_i$, we need to find the volume of traffic that entered queue i between two arbitrary points in time x and y ($x \in [0, \mathrm{T}]$, $y \in [0, \mathrm{T}]$, and $x < y$), i.e., x and y might not coincide with any t_n for $n \in \{1, \ldots, N\}$.

This volume of traffic, denoted as $V_i(x, y)$, is obtained by integrating $q^{\text{in}}_{i,n}$ over $[x, y]$ and is defined in (3.2) where m and w are the index of the time intervals s.t. $t_m \leq x < t_{m+1}$ and $t_w \leq y < t_{w+1}$. Because the QTM dynamics are *piecewise linear*, $q^{\text{in}}_{i,n}$ is a step function w.r.t. time and this integral reduces to the sum of $q^{\text{in}}_{i,n}$ over the intervals contained in $[x, y]$ and the appropriate fraction of $q^{\text{in}}_{i,m}$ and $q^{\text{in}}_{i,w}$ representing the misaligned beginning and end of $[x, y]$.

$$V_i(x, y) = (t_{m+1} - x) \frac{q^{\text{in}}_{i,m}}{\Delta t_m} + \left(\sum_{k=m+1}^{w-1} q^{\text{in}}_{i,k} \right) + (y - t_w) \frac{q^{\text{in}}_{i,w}}{\Delta t_w} \tag{3.2}$$

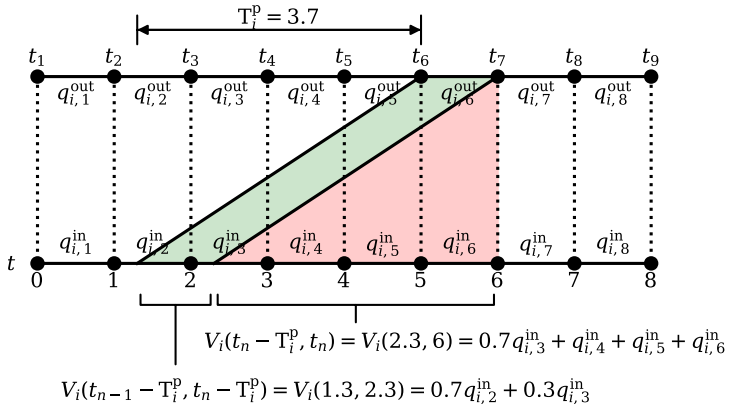


Figure 3.2: An example of using $V_i(x, y)$ for a queue i, to find the volume of traffic arriving at the stop line during interval 6, and the total volume of traffic within the link.

Using these helper variables, (Q5) represents the flow conservation principle for queue i where $V_i(t_{n-1} - T_i^p, t_n - T_i^p)$ is the volume of vehicles that reached the stop line during Δt_n. Since $\vec{\Delta t}$ and T_i^p for all queues are known a priori, the indexes m and w used by V_i can be precomputed in order to encode (3.2); moreover, (Q5) represents a non-first order Markovian update because the update considers the previous $w - m$ time steps. To ensure that the total volume of traffic traversing i (i.e., $V_i(t_n - T_i^p, t_n)$) and waiting at the stop line does not exceed the capacity of the queue, we apply (Q6). When queue i is full, $q_{i,n}^{in} = 0$ by (Q6), which forces $f_{j,i,n}$ to 0 in (Q3) and (Q4). This in turn allows the queue in i to spill back into the upstream queue j. See Fig. 3.2 for an example of the volume functions.

$$q_{i,n} = q_{i,n-1} - q_{i,n-1}^{out} + V_i(t_{n-1} - T_i^p, t_n - T_i^p) \tag{Q5}$$

$$q_{i,n} \leq Q_i - V_i(t_n - T_i^p, t_n) \tag{Q6}$$

QTM uses the objective function (O6) to minimize total delay in the network. By maximizing each flow, $f_{i,n}^{out}$, $f_{i,n}^{in}$ and $f_{i,j,n}$ against its upper bound, weighted by the time remaining until the end of the problem horizon T, the optimizer is forced to allow as much traffic volume as possible into the network and move traffic to the outside of the network as soon as possible.

$$\begin{aligned} \max & \sum_{n=1}^{N} \sum_{i=1}^{|\mathcal{Q}|} (T - t_n) \Delta t_n f_{i,n}^{out} + \sum_{n=1}^{N} \sum_{i=1}^{|\mathcal{Q}|} (T - t_n) \Delta t_n f_{i,n}^{in} \\ & + \beta \sum_{n=1}^{N} \sum_{i=1}^{|\mathcal{Q}|} (T - t_n) \sum_{j=1}^{|\mathcal{Q}|} \Delta t_n f_{i,j,n} \end{aligned} \tag{O6}$$

The first term of objective (O6) corresponds to minimizing delay. To see this, consider the objective transformation $\max J = -\min -J$ and, after expanding, we have the equivalent minimization (3.3). Without loss of generality, if we consider the case where all traffic clears the network, then the sum $\sum_n \sum_i T \Delta t_n f_{i,n}^{out}$ is a constant and (3.3) can be reduced to (3.4).

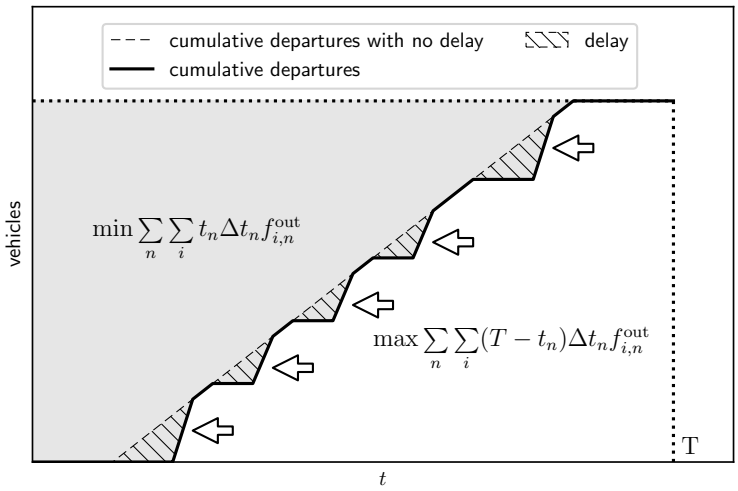


Figure 3.3: Example of a total cumulative departure curve for a network. Total delay in vehicle seconds is represented by the hatched region between the departure curve and its upper bound, the departure curve with no delay. The first term of the objective function (O6) seeks to maximize the area under the departure curve. This is equivalent to minimizing the area *above* the departure curve (shaded region) and therefore (O6) minimizes delay.

Since Δt_n is implicitly 1 in CTM, (3.4) is equal to the objective function given in [Lin and Wang, 2004], which was shown to be the minimization of total delay.

Fig. 3.3 provides a graphical interpretation of (3.3) as the difference of areas, where the first term of (O6) is the area below the cumulative departure curve for the network, the sum $\sum_n \sum_i T \Delta t_n f^{\text{out}}_{i,n}$ is the area of the dotted rectangle enclosing the curve, and the shaded area above the curve is given by (3.4).

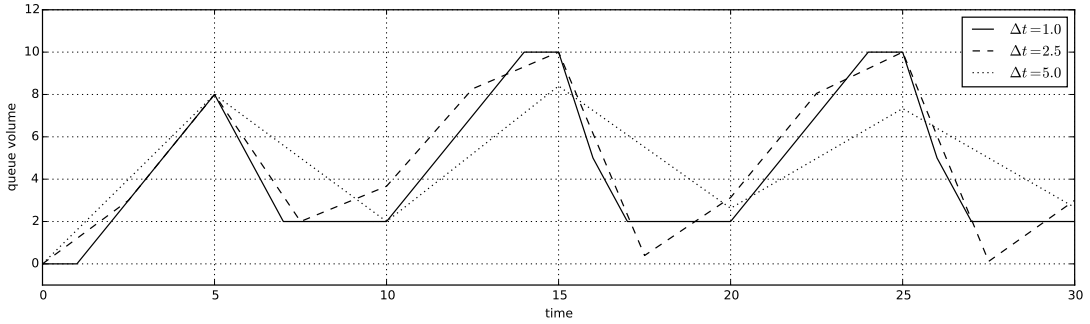
$$-\min \sum_{n=1}^{\mathrm{N}} \sum_{i=1}^{|\mathcal{Q}|} t_n \Delta t_n f^{\text{out}}_{i,n} - \sum_{n=1}^{\mathrm{N}} \sum_{i=1}^{|\mathcal{Q}|} T \Delta t_n f^{\text{out}}_{i,n} \tag{3.3}$$

$$\min \sum_{n=1}^{\mathrm{N}} \sum_{i=1}^{|\mathcal{Q}|} t_n \Delta t_n f^{\text{out}}_{i,n} \tag{3.4}$$

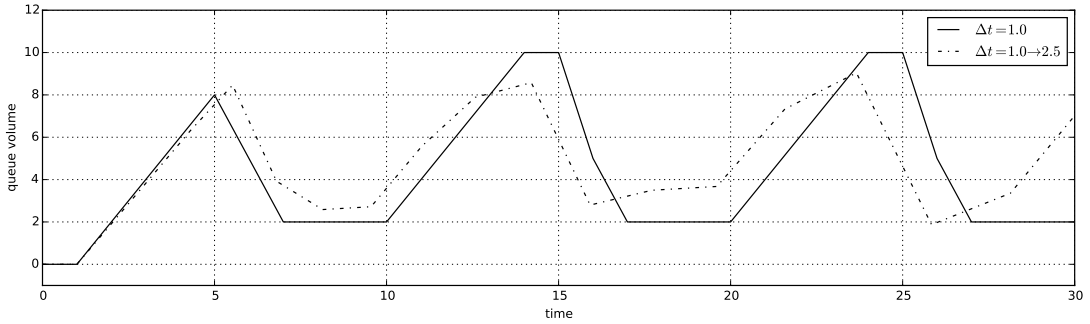
The second term of (O6) maximizes the inflow at the rate given by $\mathbf{I}$, but allows for elasticity in the case of any queue spill back that blocks an input.

The third term of (O6) ensures that the optimizer always moves vehicles from i to j when the associated traffic phase is active and j is not full. As described in [Lin and Wang, 2004], the value of β should be sufficiently small to avoid interfering with the main objective by giving too much priority to the internal flows.

To illustrate the representation tradeoff offered by non-homogeneous time intervals, we computed flows and queue volumes for a fixed signal control plan derived for homogeneous Δt_n = 1s (ground truth) using different discretizations. Fig. 3.4(a) shows the approximation of the ground truth using homogeneous Δt = 2.5 and Δt = 5.0, and Fig. 3.4(b) using non-



(a)



(b)

Figure 3.4: Approximations of a queue volume obtained using homogeneous $\vec{\Delta t} = \{1.0, \ldots, 1.0\}$ using: (a) homogeneous $\vec{\Delta t} = \{2.5, \ldots, 2.5\}$ and $\vec{\Delta t} = \{5.0, \ldots, 5.0\}$; and (b) non-homogeneous $\vec{\Delta t} = \{1.0, 1.05, 1.1, 1.16, \ldots, 2.29, 2.41, 2.5\}$ where $\Delta t_n \approx 0.0956n + 0.9044$ for $n \in \{1, \ldots, 17\}$. Here we see that (b) achieves accuracy in the near-term that somewhat degrades over the long-term, where accuracy will be less critical for receding horizon control.

homogeneous time intervals that linearly increases from 1s to 2.5s, i.e., $\Delta t_n \approx 0.0956n + 0.9044$ for $n \in \{1, \ldots, 17\}$. As Fig. 3.4(a) shows, large time steps can be rough approximations of the ground truth. Non-homogeneous discretization (Fig. 3.4(b)) exploit this fact to provide a good approximation in the initial time steps and progressively decrease precision for points far in the future.

3.3 Traffic Control with QTM encoded as a MILP

In this section, we remove the assumption that a valid control plan for all traffic lights is given and extend the LP (O6, Q1–Q6) to an Mixed-Integer LP (MILP) that also computes the optimal control plan. Formally, for all $\ell \in \mathcal{L}$, $k \in \mathcal{P}_\ell$, and interval $n \in \{1, \ldots, N\}$, the phase activation parameter $p_{\ell,k,n} \in \{0, 1\}$ becomes a free variable to be optimized. In order to obtain a valid control plan, we enforce that one phase of traffic light ℓ is always active at any interval n (Q7), and ensure cyclic phase polices where phase changes follow a fixed ordered sequence (Q8), i.e., if phase k was active during interval $n-1$ and has become inactive in interval n, then phase $k+1$ must be active in interval n. (Q8) assumes that $k+1$ equals 1 if $k = |\mathcal{P}_\ell|$.

$$\sum_{k=1}^{|\mathcal{P}_\ell|} p_{\ell,k,n} = 1 \tag{Q7}$$

$$p_{\ell,k,n-1} \leq p_{\ell,k,n} + p_{\ell,k+1,n} \tag{Q8}$$

Next, we enforce the minimum and maximum phase durations (i.e., $\Phi^{\min}_{\ell,k}$ and $\Phi^{\max}_{\ell,k}$) for each phase $k \in \mathcal{P}_\ell$ of traffic light ℓ. To encode these constraints, we use the helper variable $d_{\ell,k,n} \in [0, \Phi^{\max}_{\ell,k}]$, defined by constraints (Q9–Q13), that: (i) holds the elapsed time since the start of phase k when $p_{\ell,k,n}$ is active (Q9,Q10); (ii) is constant and holds the duration of the last phase until the next activation when $p_{\ell,k,n}$ is inactive (Q11,Q12); and (iii) is restarted when phase k changes from inactive to active (Q13). Notice that (Q9–Q13) employs the *big-M* method to turn the cases that should not be active into subsumed constraints based on the value of $p_{\ell,k,n}$. We use $\Phi^{\max}_{\ell,k}$ as our large constant since $d_{\ell,k,n} \leq \Phi^{\max}_{\ell,k}$ and $\Delta t_n \leq \Phi^{\max}_{\ell,k}$. Similarly, constraint (Q14) ensures the minimum phase time of k and is not enforced while k is still active. Figs. 3.5(a) to 3.5(c) present an example of how (Q9–Q14) work together as a function of the time n for $d_{\ell,k,n}$; the domain constraint $0 \leq d_{\ell,k,n} \leq \Phi^{\max}_{\ell,k}$ for all $n \in \{1, \ldots, \mathrm{N}\}$ is omitted for clarity.

$$d_{\ell,k,n} \leq d_{\ell,k,n-1} + \Delta t_{n-1} p_{\ell,k,n-1} + \Phi^{\max}_{\ell,k}(1 - p_{\ell,k,n-1}) \tag{Q9}$$

$$d_{\ell,k,n} \geq d_{\ell,k,n-1} + \Delta t_{n-1} p_{\ell,k,n-1} - \Phi^{\max}_{\ell,k}(1 - p_{\ell,k,n-1}) \tag{Q10}$$

$$d_{\ell,k,n} \leq d_{\ell,k,n-1} + \Phi^{\max}_{\ell,k} p_{\ell,k,n-1} \tag{Q11}$$

$$d_{\ell,k,n} \geq d_{\ell,k,n-1} - \Phi^{\max}_{\ell,k} p_{\ell,k,n} \tag{Q12}$$

$$d_{\ell,k,n} \leq \Phi^{\max}_{\ell,k}(1 - p_{\ell,k,n} + p_{\ell,k,n-1}) \tag{Q13}$$

$$d_{\ell,k,n} \geq \Phi^{\min}_{\ell,k}(1 - p_{\ell,k,n}) \tag{Q14}$$

Lastly, we constrain the sum of all the phase durations for light ℓ to be within the cycle

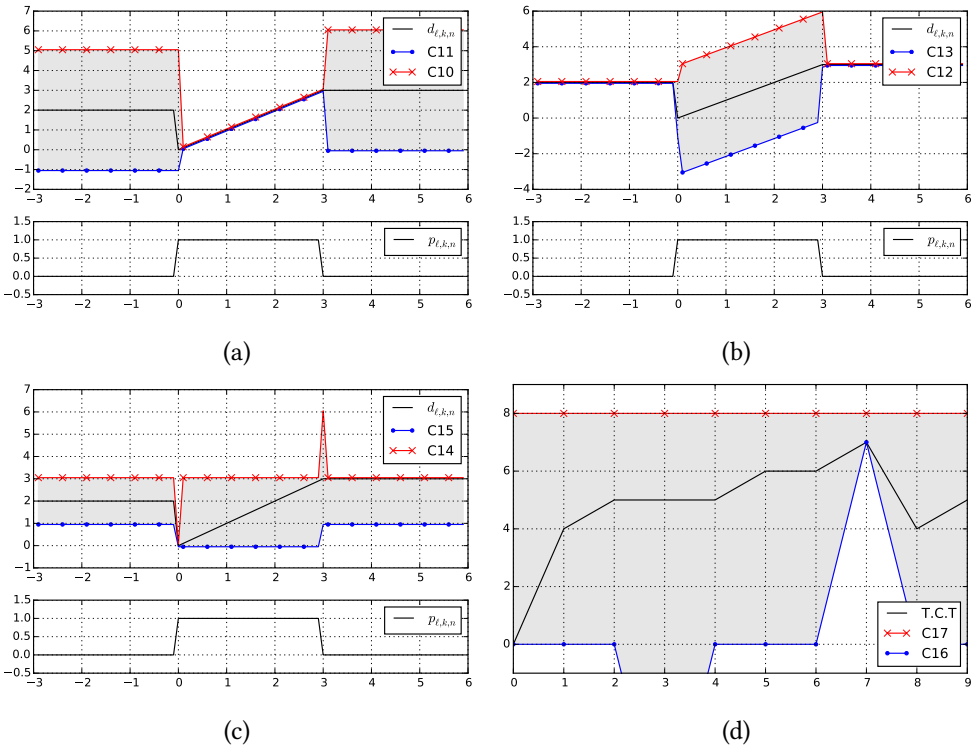


Figure 3.5: Visualization of constraints (Q9–Q16) for a traffic light ℓ as a function of time. (a–c) present, pairwise, the constraints (Q9–Q14) for phase k ($d_{\ell,k,n}$ as the black line) and the activation variable $p_{\ell,k,n}$ in the small plot. (d) presents the constraints for the cycle time of ℓ (Q15 and Q16), where T.C.T. is the total cycle time and is the left hand side of both constraints. For this example, $\Phi^{\min}_{\ell,k} = 1$, $\Phi^{\max}_{\ell,k} = 3$, $\Psi^{\min}_{\ell} = 7$, and $\Psi^{\max}_{\ell} = 8$.

time limits $\Psi_\ell^{\min}$ (Q15) and $\Psi_\ell^{\max}$ (Q16). In both (Q15) and (Q16), we use the duration of phase 1 of ℓ from the previous interval $n-1$ instead of the current interval n because (Q13) forces $d_{\ell,1,n}$ to be 0 at the beginning of each cycle; however, from the previous end of phase 1 until $n-1$, $d_{\ell,1,n-1}$ holds the correct elapse time of phase 1. Additionally, (Q15) is enforced right after the end of the each cycle, i.e., when its first phase is changed from inactive to active. The value (Q15) and (Q16) over time for a traffic light ℓ is illustrated in Fig. 3.5(d).

$$d_{\ell,1,n-1} + \sum_{k=2}^{|\mathcal{P}_\ell|} d_{\ell,k,n} \geq \Psi_\ell^{\min}(p_{\ell,1,n} - p_{\ell,1,n-1}) \tag{Q15}$$

$$d_{\ell,1,n-1} + \sum_{k=2}^{|\mathcal{P}_\ell|} d_{\ell,k,n} \leq \Psi_\ell^{\max} \tag{Q16}$$

The MILP that encodes the problem of finding the optimal traffic control plan in a QTM network is defined by (O6, Q1–Q16).

3.4 Empirical Evaluation

In this section we compare the solutions for traffic networks modeled as a QTM using homogeneous and non-homogeneous time intervals with respect to two evaluation criteria: the quality of the solution and convergence to the optimal solution vs. the number of time steps. Specifically, we compare the quality of solutions based on the total travel time and we also consider the third quartile and maximum of the observed delay distribution. The hypotheses we wish to evaluate in this paper are: (i) the quality of the non-homogeneous solutions is at least as good as the homogeneous ones when the number of time intervals N is fixed; and (ii) the non-homogeneous approach requires less time intervals (i.e., smaller N) than the homogeneous approach to converge to the optimal solution. In the remainder of this section, we present the traffic networks considered in the experiments, our methodology, and the results.

3.4.1 Networks

We consider three networks of increasing complexity (Fig. 3.6): an avenue crossed by three side streets; a 2-by-3 grid; and a 3-by-3 grid with a diagonal avenue. The queues receiving vehicles from outside of the network are marked in Fig. 3.6 and we refer to them as input queues. The maximum queue capacity (Q_i) is 60 vehicles for non-input queues and infinity for input queues to prevent interruption of the input demand due to spill back from the stop line. The traversal time of each queue i (T_i^p) is set at 9s (a distance of 125m with a free flow speed of 50km/h). For each street, flows are defined from the head of each queue i into the tail of the next queue j; there is no turning traffic ($Pr_{i,j} = 1$), and the maximum flow rate between queues, $F_{i,j}$, is set at 5 vehicles/s. All traffic lights have two phases, north-south and east-west, and lights 2, 4 and 6 of network 3 have the additional northeast-southwest phase to control the diagonal avenue. For networks 1 and 2, $\Phi_{\ell,k}^{\min}$ is 1s, $\Phi_{\ell,k}^{\max}$ is 3s, $\Psi_\ell^{\min}$ is 2s, and $\Psi_\ell^{\max}$ is 6s, for all traffic light ℓ and phase k. For network 3, $\Phi_{\ell,k}^{\min}$ is 1s and $\Phi_{\ell,k}^{\max}$ is 6s for all ℓ and k; and $\Psi_\ell^{\min}$ is 2s and $\Psi_\ell^{\max}$ is 12s for all lights ℓ except for lights 2, 4 and 6 (i.e., lights also used by the diagonal avenue) in which $\Psi_\ell^{\min}$ is 3s and $\Psi_\ell^{\max}$ is 18s.

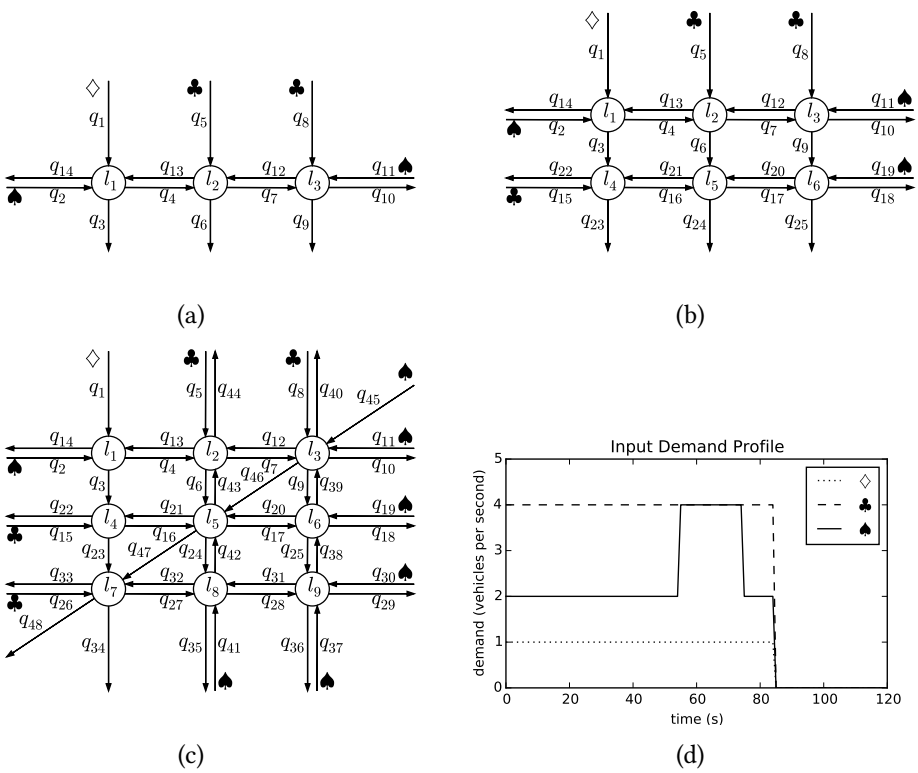
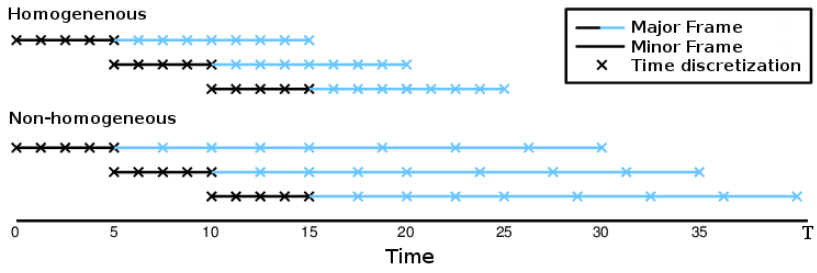
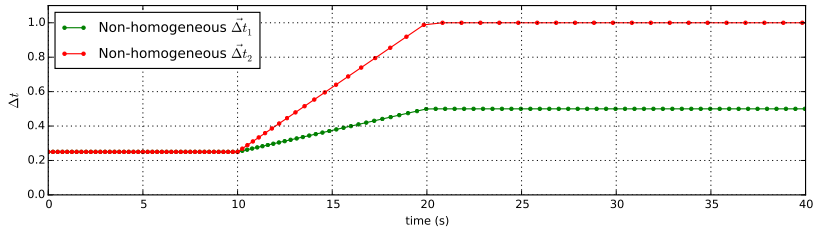


Figure 3.6: (a–c) Networks used to evaluate the QTM performance. (d) Demand profile of the queues marked as ♢, ♣, and ♠ for our experiments.



(a)



(b)

Figure 3.7: (a) Receding horizon control. In this example, the problem horizon T is 40s. The major frames for MILP optimization are discretized in 12 time intervals (N = 12) and they span 15s and 30s for homogeneous and non-homogeneous discretizations, respectively. The minor frames represent the prefix of the major frame MILP optimization that is executed. The horizon recedes by the minor frame duration after each execution. (b) The two non-homogeneous discretizations used in the experiments, shown here with a major frame duration of 40s. From the end of the minor frame time, Δt is linearly interpolated over 10s, from 0.25 to 0.5 for Non-homogeneous $\vec{\Delta t}_1$, and 0.25 to 1.0 for Non-homogeneous $\vec{\Delta t}_2$. Δt is then held constant to the end of the major frame time. The number of intervals in $\vec{\Delta t}_1$ is N = 80, and in $\vec{\Delta t}_2$ is N = 68.

3.4.2 Experimental Methodology

For each network, a constant background level traffic is injected in the network in the first 55s to allow the solver to settle on a stable policy. Then a spike in demand is introduced in the queues marked as ♠ (Fig. 3.6) from time 55s to 70s to trigger a policy change. From time 70s to 85s, the demand is returned to the background level, and then reduced to zero for all input queues. We extend the problem horizon T until all vehicles have left the network. By clearing the network, we can easily measure the total travel time for all the traffic as the area between the cumulative arrival and departure curves measured at the boundaries of the network. The background level for the input queues are 1, 4 and 2 vehicles/s for queues marked as ♢, ♣ and ♠ (Fig. 3.6(d)), respectively; and during the high demand period, the queues ♠ receive 4 vehicles/s.

For both homogeneous and non-homogeneous intervals, we use the MILP QTM formulation in a receding horizon manner: a control plan is computed for a pre-defined horizon (smaller than T) and only a prefix of this plan is executed before generating a new control plan. Fig. 3.7(a) depicts our receding horizon approach and we refer to the planning horizon as a major frame and its executable prefix as a minor frame. Notice that, while the plan for a minor frame is being executed, we can start computing the solution for the next major frame

based on a forecast model.

To perform a fair comparison between the homogeneous and non-homogeneous discretizations, we fix the size of all minor frames to 10s and force it to be discretized in homogeneous intervals of 0.25s. For the homogeneous experiments, Δt is kept at 0.25s throughout the major frame; therefore, given N, the major frame size equals N/4 seconds for the homogeneous approach. For the non-homogeneous experiments, we increase Δt linearly from the end of the minor frame for 10s and then hold it constant to the end of the major frame. We use two discretizations as shown in Fig. 3.7(b): Non-homogeneous $\vec{\Delta t}_1$ from 0.25 to 0.5, and Non-homogeneous $\vec{\Delta t}_2$ from 0.25 to 1.0. For a given $N > 40$, the major frame size used by this non-homogeneous approach is $10.375 + 1.25(N - 40)$ seconds for $\vec{\Delta t}_1$, and $10.375 + 0.625(N - 40)$ seconds for $\vec{\Delta t}_2$. Once we have generated a series of minor frames, we concatenate them into a single plan and compute the flow through the network using the QTM LP formulation with a fixed (homogeneous) Δt of 0.25s. We also compare both receding horizon approaches against the optimal solution obtained by computing a single control plan for the entire control horizon (i.e., $[0, T]$) using a fixed Δt of 0.25s.

For all our experiments, we used Gurobi^{TM} as the MILP solver with 12 threads on a 3.1GHz AMD Opteron^{TM} 4334 processor with 12 cores. We limit the MIP gap accuracy to 0.1% and the time cutoff for solving a major frame to 3000s for the receding horizon approaches and unbounded in order to determine the optimal minimum travel time solution to which all other solutions are compared. All our results are averaged over five runs to account for Gurobi's stochastic strategies.

3.4.3 Results

Figs. 3.8(a), 3.8(c) and 3.8(e) show, for each network, the increase in the total travel time w.r.t. the optimal solution as a function of N. As we hypothesized, the non-homogeneous discretizations requires less time intervals (i.e., smaller N) to obtain a solution with the same total travel time, and $\vec{\Delta t}_2$ converges before $\vec{\Delta t}_1$. This is important because the size of the MILP, including the number of binary variables, scales linearly with N; therefore, the non-homogeneous approach can scale up better than the homogeneous one (e.g., Fig. 3.8(e)). Also, for homogeneous and non-homogeneous discretizations, finding the optimal solution of major frames with large N might require more time than our imposed 3000s time cutoff and, in this case, Gurobi returns a feasible control plan that is far from optimal. The effect in the total travel time of these poor solutions can be seen in Fig. 3.8(e) for $N > 120$.

The distribution of the total delay observed by each vehicle while traversing the network is shown in Figs. 3.8(b), 3.8(d) and 3.8(f). Each group of box plots represents a different value of N: when the non-homogeneous $\vec{\Delta t}_2$ first converges; when the homogeneous Δt first converges; and the final solution itself. In all networks, the quality of the solutions obtained using both of the $\vec{\Delta t}_1$ and $\vec{\Delta t}_2$ and is better or equal than using homogeneous Δt for fixed N in both the total travel time and *fairness*, i.e., smaller third quartile and maximum delay.

To further illustrate the differences between homogeneous and non-homogeneous discretizations, Fig. 3.9 shows the cumulative arrival and departure curves and the how delay evolves over time for q_1 of network 2 (Fig. 3.6(b)). In Fig. 3.9(a), the comparison is done when non-homogeneous $\vec{\Delta t}_2$ first converges (i.e., point I in Fig. 3.8(c)) and for this value of N, the

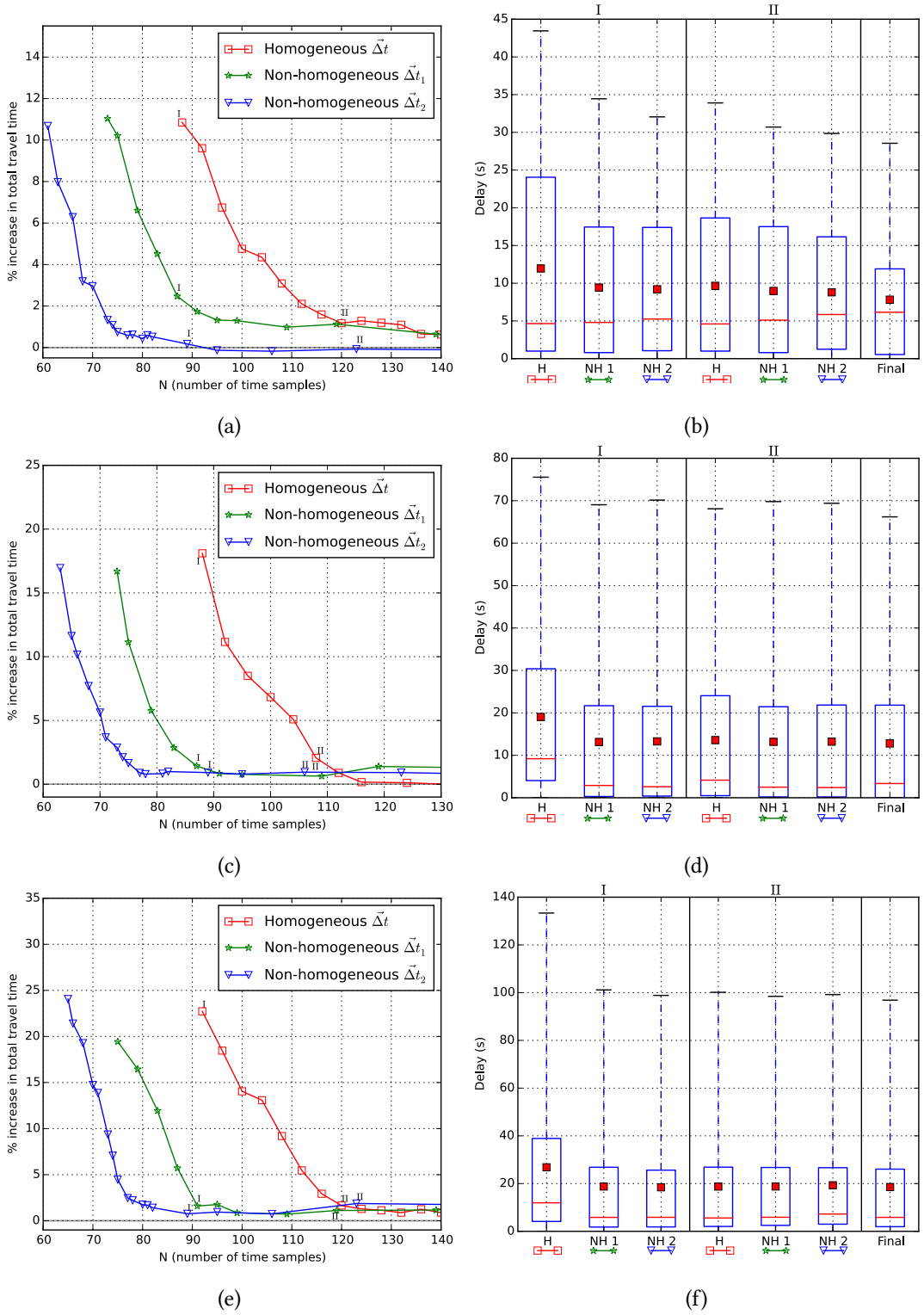
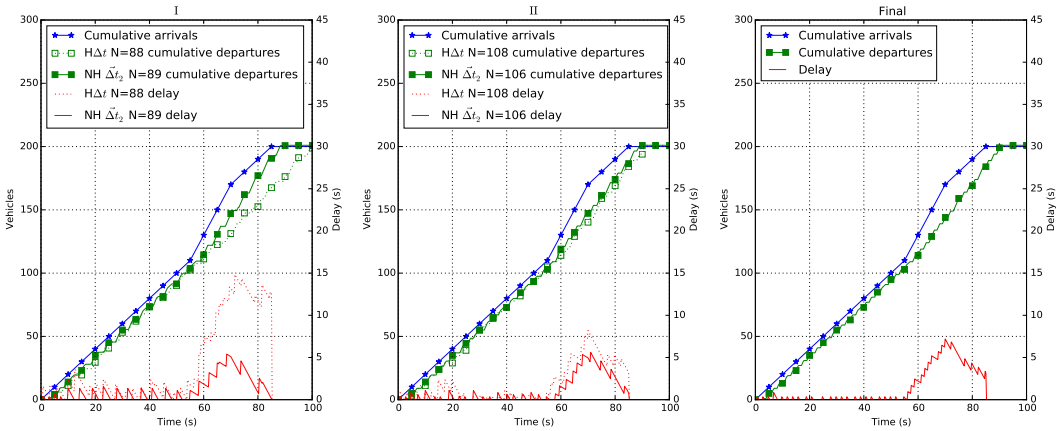


Figure 3.8: Increase in the total travel time w.r.t. the optimal solution as a function of N (a,c,e) and distribution of the total delay of each car for different values of N (b,d,f). For each row, the Roman numeral on top of the box plots corresponds to points on the travel time plot marked with the same numeral. The mean of the total delay is presented as a red square in the box plots. Plots in the i-th row correspond to the results for the i-th network in Fig. 3.6. Non-homogeneous (NH) achieves much better solutions at smaller N than Homogeneous (H).



(a) (b) (c)

Figure 3.9: Cumulative arrival and departure curves and delay for queue 1 in the 2-by-3 network (Fig. 3.6(b)). The labels on top of each plot match the labels in Figs. 3.8(c) and 3.8(d). (c) presents the same curves for the optimal solution. Non-homogeneous (NH $\vec{\Delta}t_2$) provides near-optimal signal plans over a longer time horizon than Homogeneous (H) when the number of time intervals N is small.

major frame size in seconds of the non-homogeneous approach is 19.125s longer than the homogeneous one. This allows the MILP solver to "see" 19s further in the future when using non-homogeneous discretization and find a coordinated signal policy along the avenue to dissipate the extra traffic that arrives at time 55s. The shorter major frame of the homogeneous discretization does not allow the solver to adapt this far in advance and its delay observed after 55s is much larger than the non-homogeneous one. Once the homogeneous Δt has converged (Fig. 3.9(b)), it is also able to anticipate the increased demand and adapt well in advance and both approaches generate solutions close to optimum (Fig. 3.9(c)).

3.5 Summary

In this chapter, we showed how to formulate a novel queue transmission model (QTM) of traffic flow with non-homogeneous time steps as a linear program. We then proceeded to allow the traffic signals to become discrete variables subject to a delay minimizing optimization objective and standard traffic signal constraints leading to a final MILP formulation of traffic signal control with non-homogeneous time steps. We experimented with this novel QTM-based MILP control in a range of traffic networks and demonstrated that the non-homogeneous MILP formulation achieved (i) substantially lower delay solutions, (ii) improved per-vehicle delay distributions, and (iii) more optimal travel times over a longer horizon in comparison to the homogeneous MILP formulation with the same number of binary and continuous variables. Altogether, this work represents a major step forward in the scalability of MILP-based jointly optimized traffic signal control via the use of a non-homogeneous time traffic models and thus helps pave the way for fully optimized joint urban traffic signal controllers as an improved successor technology to existing signal control methods.

Chapter 4

QTM Extensions and Microsimulation Validation

4.1 Overview

Many cities are increasingly looking to public transit options such as light rail that are less expensive and often more reliable than heavy rail in order to reduce the number of conventional traffic commuters [Thompson, 2003]. Since light rail often operates at street-level with exclusive right-of-way and requires coordination with conventional traffic networks and signal control, a major concern in light rail installation is whether enough commuters will switch to it to offset the additional constraints it places on traffic signal control.

Unfortunately, many large cities still use some degree of *fixed-time* control [El-Tantawy et al., 2013] even if they also use *actuated* or *adaptive* control methods such as SCATS [Sims and Dobinson, 1980] or SCOOT [Hunt et al., 1981]; while these methods may support signal pre-emption for light rail crossing, they are unable to autonomously adapt the signal plan to the light rail schedule, hence posing problems for their effective integration with conventional traffic signal control. A more recent trend in the traffic signal control literature proposes the use of *optimized* controllers (that incorporate elements of both adaptive and actuated control) as evidenced in a variety of approaches including mixed integer linear programming (MILPs) [Lo, 1998; Gartner et al., 1974; Gartner and Stamatiadis, 2002; Lin and Wang, 2004; He et al., 2011; Han et al., 2012, 2016a; Guilliard et al., 2016; Wada et al., 2017], heuristic search [Lo et al., 1999; He et al., 2010], queuing delay optimization [Varaiya, 2013; Li and Zhang, 2014], scheduling-driven control [Xie et al., 2012; Smith et al., 2013], and reinforcement learning [El-Tantawy et al., 2013]. While these approaches hold out the promise of more highly optimized traffic control methods, to date, *none* have studied the optimal integration of light rail schedule constraints with their respective methods nor the impact that such integration would have on traffic delay.

Nonetheless, the sub-optimal integration of traffic signal optimization and light rail schedules has been done before, such as in [Stevanovic et al., 2008], which uses a genetic algorithm coupled with a microsimulator to optimize a subset of traffic signal and transit priority request parameters. Due to the usage of a genetic algorithm, this approach does not necessarily find the global optimum. Another example of signal plan optimization taking into account a schedule of transit priority requests is [He et al., 2014] which represents this problem as a MILP that minimizes a multi-modal delay objective. Additionally, this approach incorporates virtual priority requests to represent vehicle platoon arrivals in order to improve coordination between intersections. This approach is also not globally optimal since each intersection is solved separately. Moreover, while the model includes queuing delay and clearance times, it does not consider start up lost time. Another non-optimal MILP-based approach is introduced in [Christofa et al., 2016] and their MILP represents platoon-based flows to optimize traffic signals with transit priority requests, but not the signal preemption associated with rail transit. The obtained MILP is then solved progressively on pairs of intersections along an arterial route and does not necessarily find the global optimum.

To address the deficiencies of these sub-optimal approaches, we introduce a MILP-based algorithm for optimizing traffic signals constrained by light rail schedules in which it is computationally feasible to find the optimal solution. To the best of our knowledge, this is the first algorithm capable of finding the optimal integration of light rail schedule constraints. In order to do so, we leverage the *Queue Transmission Model* (QTM) from Chapter 3, a MILP

model of traffic signal optimization where expected traffic queues and flows are continuous variables, traffic signals are discrete variables, and the overall optimization objective is to minimize delay. Among alternative MILP-based control methods cited previously, the advantages of QTM are improved scalability through the use of non-homogeneous time steps as demonstrated in Chapter 3, and a focus on the accurate modeling of travel delay between intersections critical for prioritizing light rail arrivals.

Our approach to modelling light rail schedules also allow us to compute optimal fixed-time control policies with light rail constraints and is capable of finding the optimal splits, offsets and cycle time. This is an improvement on previous approaches to this problem (e.g., [Lo, 1998; Lin and Wang, 2004; Wada et al., 2017]) which require the cycle time to be fixed exogenously (i.e., a parameter of the algorithm) and are only able to optimize the splits and offsets. Computing the optimal cycle times has a impact beyond fixed-time control since adaptive controllers (e.g., [Sims and Dobinson, 1980]) also require cycle times to be known a priori and our method can be directly used there.

We make the follow key contributions:

1. The first method to globally optimize traffic signals integrated with light rail schedule constraints.

2. We provide a novel fixed-time controller to optimize cycle times, phase splits and offsets. The fixed-time control schedules can include light rail schedule constraints and common cycle length constraints, and can be incorporated *immediately* into existing fixed-time traffic controller infrastructure.

3. We provide a novel way to model lost time directly as a signal timing constraint and we show that it is critical to finding optimized signal plans.

4. We run a comprehensive suit of experiments using a microsimulator to validate the effectiveness of these contributions, both quantitatively and through visual inspection of the simulation results. Gaining insights into the optimal solution's properties can also help to further improve existing control strategies and provides a benchmark.

Our experiments show that optimal adaptive control can reduce traffic delay by up to 58.7% over optimal fixed-time control when light rail is introduced, and virtually nullifies its impact when compared to using fixed-time control before the introduction of light rail. Ultimately, these results demonstrate a win-win situation where both vehicle traffic and light rail commuters benefit through the application of MILP-based optimization to jointly manage both light rail schedule priority and traffic networks.

4.2 QTM Extensions

To investigate the impact of light rail schedules on conventional traffic networks we need a model of both traffic flow and light rail constraints. As a model of traffic flow, we leverage the Queue Transmission Model and the MILP (O6, Q1–Q16) which encodes the problem of finding the optimized adaptive traffic control plan in a QTM network without light rail.

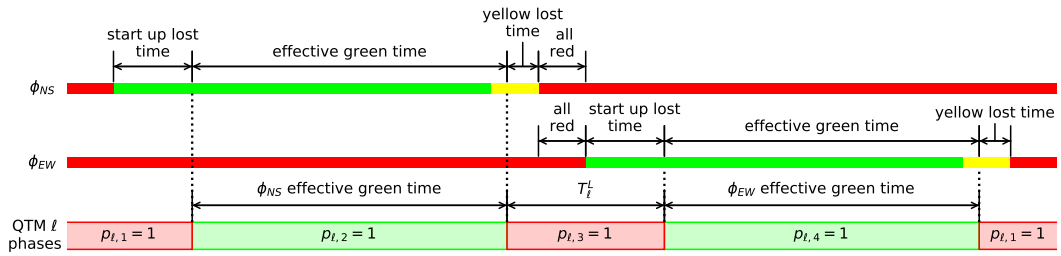


Figure 4.1: Signal timing for a crossroad intersection with two phases ϕ_{NS} and ϕ_{EW} controlling the north-south and east-west approaches respectively. The intersection is modeled using a QTM light ℓ, with four consecutive phases: $\mathcal{P}_\ell = \{p_{\ell,1}, p_{\ell,2}, p_{\ell,3}, p_{\ell,4}\}$ with $\mathcal{Q}^{\mathcal{P}}_{NS} = \{p_{\ell,2}\}$, $\mathcal{Q}^{\mathcal{P}}_{EW} = \{p_{\ell,4}\}$ and $p_{\ell,1}, p_{\ell,3}$ are the lost time phases of fixed duration T^L_ℓ. The active states of $p_{\ell,2}$ and $p_{\ell,4}$ represent the *effective* green time of their corresponding approaches, where traffic flows at the free flow speed. The lost time is represented by the active states of $p_{\ell,1}$ and $p_{\ell,3}$, which inhibit all traffic flow, forcing the solver to clear the intersection between each signal change and incur lost time delay.

4.2.1 Lost Time

An additional delay experienced by traffic within a signalized network is the time lost during signal changes. This lost time is made up of several components:

- **Start up lost time**: the time require for a driver waiting at the stop line to react to a green signal and accelerate up to the free flow speed.
- **Yellow lost time**: the remaing time of a yellow signal where drivers react and come to a stop.
- **All red time**: the time preceding the start of every green signal where all approaches are held red to allow vehicles from the previous phase to clear the intersection.

As with other MILP formulations, the active states of QTM phase variables represent the *effective* green time of the associated traffic signal phases, i.e, the time during which vehicles are flowing at the free flow speed, but switch instantaneously between phases in the cycle, without consideration for the lost time associated with the signal change [Webster, 1958; Wolshon and Pande, 2016]. We extend QTM to model lost time by inserting additional fixed duration phase variables into the cycle at each signal change that inhibit the flow of traffic when active. If the lost time per signal change for light ℓ is T^L_ℓ, then we fix the duration of lost time phase k with $\Phi^{\min}_{\ell,k} = \Phi^{\max}_{\ell,k} = T^L_\ell$, and the solver cannot transition from one signal state to the next without first incurring a delay of T^L_ℓ. In general, a total of n additional phase variables are needed per cycle, where n is the number of signal phases in the cycle.

To obtain a signal plan using only the original n phases, the solution of a QTM with lost time network is post-processed by removing the lost time phases and adjusting the start (end) of each green time by the start up (yellow) lost time, leaving an all red time between signal changes. Fig. 4.1 shows a signal plan for a crossroad with two phases ϕ_{NS} and ϕ_{EW}, modeled using a QTM light ℓ with four phases. The durations of $p_{\ell,2}$ and $p_{\ell,4}$ represent the effective

green time of signal phases ϕ_{NS} and ϕ_{EW}, and are optimized by the solver. The durations of $p_{\ell,1}$ and $p_{\ell,3}$ are fixed and represent the lost time associated with the signal changes.

4.2.2 QTM as a Fixed-Time Controller

We can further extend QTM to compute an optimized control plan with fixed phase durations. For all $\ell \in \mathcal{L}$, $k \in \mathcal{P}_\ell$, we introduce the new variable $\phi^{\text{fixed}}_{\ell,k} \in [\Phi^{\min}_{\ell,k}, \Phi^{\max}_{\ell,k}]$ and replace the bounds constraints on $d_{\ell,k,n}$ (that is, $d_{\ell,k,n} \leq \Phi^{\max}_{\ell,k}$ and Q14) with fixed the duration constraints (Q17) and (Q18).

$$d_{\ell,k,n} \leq \phi^{\text{fixed}}_{\ell,k} \tag{Q17}$$

$$d_{\ell,k,n} \geq \phi^{\text{fixed}}_{\ell,k} - \Phi^{\max}_{\ell,k} p_{\ell,k,n} \tag{Q18}$$

Similarly to the variable phase duration constraints, the *big-M* method is employed in (Q17) and (Q18), using $\Phi^{\max}_{\ell,k}$ as the constant, to enforce $d_{\ell,k,n} = \phi^{\text{fixed}}_{\ell,k}$ only while the phase is inactive. The constraints (Q9–Q13, Q15, Q16, Q17, Q18), allow the fixed-time controller to optimize the phase splits, cycle length and offset for each light ℓ.

A further utility to aid coordination between intersections, is to enforce a common cycle length among a set of lights. To force a common cycle length optimized by QTM, we introduce the new variable $\psi^{\text{fixed}} \in [\max_\ell \{\Psi^{\min}_\ell\},\ \min_\ell \{\Psi^{\max}_\ell\}]$. We can then replace (Q15) and (Q16) with the new constraints (Q19) and (Q20).

$$d_{\ell,1,n-1} + \sum_{k=2}^{|\mathcal{P}_\ell|} d_{\ell,k,n} \leq \psi^{\text{fixed}} \tag{Q19}$$

$$d_{\ell,1,n-1} + \sum_{k=2}^{|\mathcal{P}_\ell|} d_{\ell,k,n} \geq \psi^{\text{fixed}} - \mathcal{M}(1 - p_{\ell,1,n} + p_{\ell,1,n-1}) \tag{Q20}$$

(Q19) and (Q20) constrain the sum of all phase durations for light ℓ to equal ψ^{fixed}. Similar to (Q15), constraint (Q20) is enforced right at the end of each cycle using the *big-M* method, where $\mathcal{M} = \max_\ell \{\Psi^{\max}_\ell\}$. We consider here a single global ψ^{fixed} for all $\ell \in \mathcal{L}$, however it would be trivial to have disjoint subsets of $\mathcal{L}$ corresponding to different regions of the network, each with its own localized ψ^{fixed}.

With the addition of constraints (Q17) to (Q20), We now have four different controllers available to us:

1. MILP (O6, Q1–Q16), a fully optimized adaptive controller.
2. MILP (O6, Q1–Q13, Q15–Q18), a fixed-time controller able to optimize the phase splits, cycle length and offset for each light ℓ.
3. MILP (O6, Q1–Q13, Q17–Q20), a fixed-time controller, but with the additional constraint of a common cycle length between lights.
4. MILP (O6, Q1–Q14, Q19, Q20) an optimized adaptive controller but also with a fixed, common cycle length.

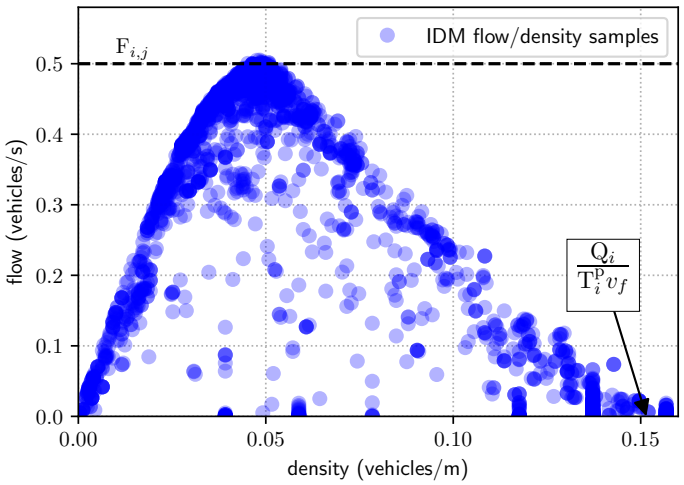


Figure 4.2: Flow-density samples from an IDM microsimulation of Network 2 along $q_{5,}$ to $q_{8,}$, showing how QTM parameters $F_{i,j} = 0.5$ and $Q_i = 60$ are calibrated to the simulation.

4.2.3 Light Rail Constraints

As a novel extension of the QTM to incorporate a fixed-schedule light rail, we add constraints to the MILP model to fix the free variable $p_{\ell,k,n}$ for all n s.t. the light rail uses phase k of ℓ at time n. Formally, given a schedule as a function $S_\ell(k,n) \in \{0,1\}$ where 1 represents that the light rail uses phase k of ℓ at time n, we replace (Q13) and (Q14) by (Q21) and (Q22) when $\sum_{k\in\mathcal{P}_\ell} S_\ell(k,n) > 0$.

$$p_{\ell,k,n} = S_\ell(k,n) \tag{Q21}$$

$$d_{\ell,k,n} = d_{\ell,k,n-1} \tag{Q22}$$

(Q21) enforces that the correct phase k is active when the light rail reaches the traffic light ℓ, and (Q22) ensures that the light rail can pass through ℓ even if more than the maximum phase time $\Phi^{\max}_{\ell,k}$ is necessary.

4.3 Empirical Evaluation

In this section we compare the solutions for traffic networks modeled using QTM with lost time before and after the introduction of a light rail. We consider both fixed-time control, i.e., a non-adaptive control plan, and optimized adaptive control obtained by solving the MILP (O6, Q1–Q16, Q21, Q22) for the optimized controller, and the MILP (O6, Q1–Q13, Q15–Q18, Q21, Q22) for the fixed controller. For comparison, we also use the optimized adaptive controller with common cycle time constraints, MILP (O6, Q1–Q14, Q19–Q22), and the fixed controller with common cycle time constraints, MILP (O6, Q1–Q13, Q17–Q22).

All the computed controllers are simulated using an Intellegent Driver Model (IDM) based microsimulator Treiber et al. [2000]. As a car following model, IDM will maintain a given safe time headway between vehicles, while also trying to achieve the given desired velocity. Ve-

hicles encountering red signals or stationary vehicles will decelerate and stop. Upon a signal change from green to yellow, the simulator will apply a given braking deceleration to any vehicles estimated not to cross the stopline within the time before the red signal, if continuing at their current speed. The simulated total travel time and observed delay distribution of each controller are used as comparison metrics. Our hypothesis is that our optimized adaptive approach is able to mitigate the impact of introducing light rail w.r.t. both metrics.

Microsimulation Parameters: We choose IDM parameters similar to those suggested in Treiber et al. [2000], that give realistic values for urban traffic with a flow capacity of 0.5 vehicles/s and a jam density of 0.15 vehicles/m. To simulate the average conditions, we give all vehicles the same parameter values: length $l = 4.67$ m, desired velocity $v_0 = 15$ m/s, safe time headway $T = 1$ s, maximum acceleration $a = 2$ m/s^2, desired deceleration $b = 3$ m/s^2, acceleration exponent $\delta = 4$, and jam distances $s_0 = s_1 = 2$ m. Fig. 4.2 shows flow-density samples from an IDM microsimulation with these values, and how the QTM flow parameters used in the experiments are calibrated.

Network Parameters: We consider two networks of differing complexity: an arterial crossed by four side streets (Fig. 4.3(a)) and a 3-by-3 grid (Fig. 4.3(b)). The queues receiving vehicles from outside of the network are marked in Fig. 4.3 and we refer to them as input queues. The maximum queue capacity (Q_i) is 60 vehicles for non-input queues and infinity for input queues to prevent interruption of the input demand due to spill back from the stop line. The free flow speed $v_f = 13.2$ m/s and the traversal time of each queue i (T_i^p) is set at 30s, except for the output queues on Network 1 where the traversal time is 10s. For each street, flows are defined from the head of each queue i into the tail of the next queue j; there is no turning traffic ($Pr_{i,j} = 1$), and the maximum flow rate between queues, $F_{i,j}$, is set at 0.5 vehicles/s. All traffic lights have two phases, north-south and east-west, and for each traffic light ℓ and phase k, $\Phi_{\ell,k}^{\min}$ is 10s, $\Phi_{\ell,k}^{\max}$ is 60s, $\Psi_{\ell}^{\min}$ is 40s, and $\Psi_{\ell}^{\max}$ is 140s. Whenever lost time is considered, we use $T_{\ell}^{L} = 10$s for all $\ell \in \mathcal{L}$, made up of 6s of startup lost time, 2s of yellow lost time and 2s of all red.

Demand Profiles: Each network is evaluated at increasing demand levels up to the point where $f_{i,n}^{\text{in}}$ becomes saturated. To simulate the effect of random arrivals, we use demand

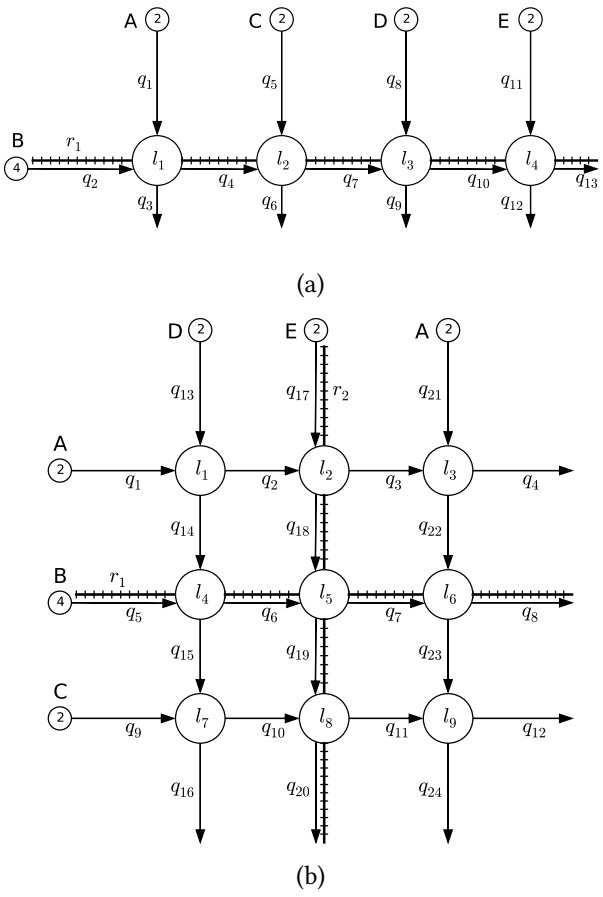


Figure 4.3: (a) Network 1, an arterial road with parallel light rail. (b) Network 2, an urban grid with crisscrossing streets and light rail.

profiles with flow rates that vary randomly every 100s for a total duration of 600s:

$$
\begin{aligned}
I_{i,n} &= \frac{1}{\Delta t_n}\max(\xi\Omega_i \mathrm{w}_i(t_n), \Omega_i) \\
w_i(t_n) &= \begin{cases} \vec{\mathrm{r}}_k & \text{if } 100(k-1) \le t_n < 100k \text{ and } 1 \le k \le 6 \\ 0 & \text{if } t_n \ge 600 \end{cases} \\
\vec{\mathrm{r}} &= \frac{1}{2} + \frac{1}{4}\vec{X} \qquad \text{where } \vec{X} \sim \text{multinominal}\left(\mathrm{v}_i, \frac{1}{6}\mathbb{1}_6\right) \\
\mathrm{v}_i &= \begin{cases} 5 & \text{if } i \text{ is labeled A} \\ 9 & \text{if } i \text{ is labeled B} \\ 7 & \text{if } i \text{ is labeled C} \\ 3 & \text{if } i \text{ is labeled D} \\ 5 & \text{if } i \text{ is labeled E} \end{cases}
\end{aligned}
$$

where:

- $\mathrm{w}_i(t_n)$ is the weight function for queue i at time t_n
- Ω_i is the maximum inflow rate in vehicles per Δt_n, as annotated at the start of queue i in Fig. 4.3
- $\xi \in (0, 2]$ is the scaling factor for the demand level being evaluated
- v_i is a constant for queue i corresponding to the letter label of i in Fig. 4.3
- multinominal $\left(\mathrm{v}_i, \vec{p_n}\right)$ returns a vector of n random integers that sum to v_i and are drawn from the multinominal distribution with v_i trials with uniform probability vector $\vec{p_n}$. Using a random vector that sums to v_i ensures that across all experiments, the total number of vehicles entering i will be the same.

An example of two different demand profiles is shown in Fig. 4.4.

Light Rail Parameters: we use two different light rail schedules: a slow light rail with a crossing duration of 50s, a period of 200s, and a travel time of 100s between lights (Fig. 4.6(a)); and a fast light rail with a crossing duration of 20s, period of 160s, and travel time of 80s between lights (Fig. 4.6(b)). On Network 2, the North-South schedule is offset by 100s for the slow light rail and 80s for the fast light rail to avoid a collision at l_5.

Evaluation: We evaluate each network in two scenarios: before the introduction of light rail and after, and in each scenario using both a fixed-time controller and an optimized adaptive controller. For each experiment, we perform one or more *runs* where a run consists of: (i) generate a random demand profile P from a multinominal distribution as described above; (ii) compute the signal plan using QTM configured as either an optimized adaptive controller or a fixed-time controller for the demand profile P; and (iii) evaluate the obtained signal plan by microsimulation on the demand profile P using IDM. We use a problem horizon T large

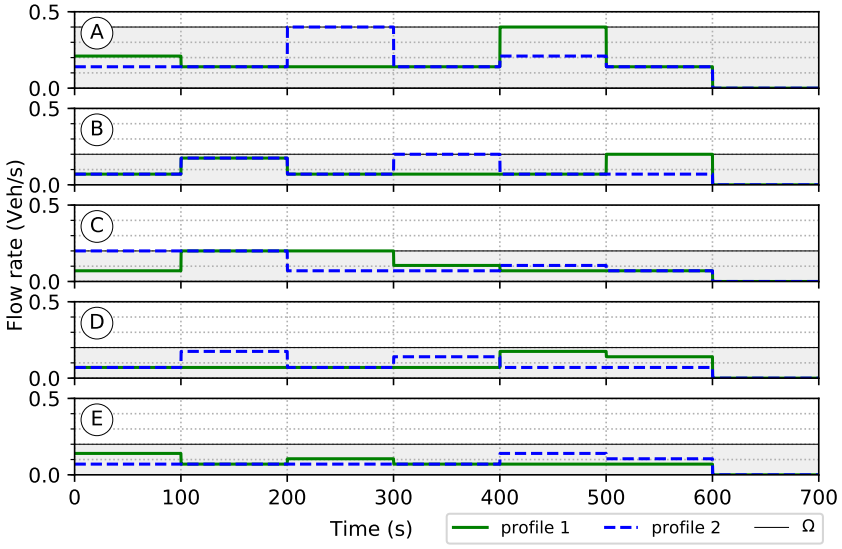
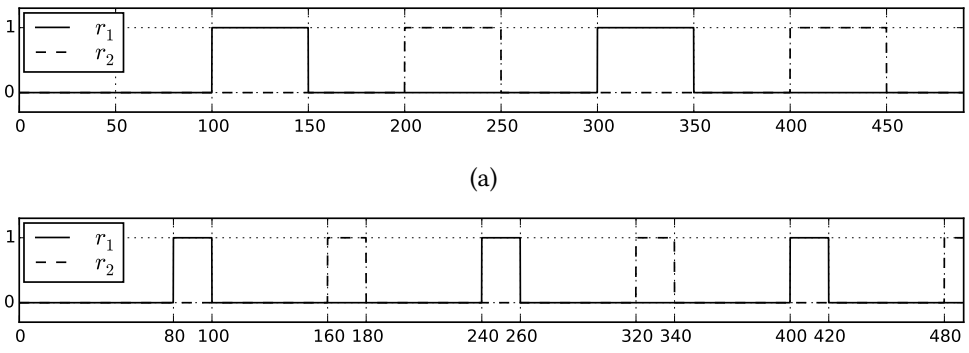


Figure 4.4

Figure 4.5: Example of two different demand profiles (inflow rates) applied to the Network 1 inputs corresponding to the letter labels A–E in each plot, where Ω is the maximum inflow rate as annotated on each input in vehicles per Δt.



(a)

(b)

Figure 4.6: Light rails schedules: (a) **slow** (long and infrequent) light rail; and (b) **fast** (short and frequent) light rail. The schedules shown are of $S_1(1, n)$ for the r_1 crossing during phase 1 at l_1 in Network 1, and of $S_4(1, n)$ for the r_1 crossing during phase 1 at l_4 in Network 2, and $S_2(2, n)$ for the r_2 crossing during phase 2 at l_2 in Network 2. The schedules will be offset by 80s at each subsequent crossing for the fast light rail, and 100s for the slow light rail.

enough, typically in the range 1000s – 1500s, to allow all traffic to clear the network, that lets us measure the incurred delay in all the vehicles.

For the experiment on Network 1, we perform 10 runs and report their average delay, and we concatenate the observed delays and number of stops of all the 10 runs for the reported boxplots and cumulative distributions. For Network 2, we report 2 different experiments: the first done in a single run (i.e., using a single random demand profile for both the controller and microsimulation), and the second using a single run in which the signal plan is further evaluated by microsimulating with 9 additional random profiles. The latter experiment allows us to evaluate the robustness of the controllers w.r.t. changes to the assumed input levels since the microsimulation will be performed using demand profiles that the controllers were not optimized for.

For lost time comparison, we use two different configurations of each controller: one with QTM incurring lost time delay and the other without. Before microsimulation, we adjust the green time of the signal plan from the lost time controller to account for the start up and yellow lost time, as illustrated in Fig. 4.1.

MILP Solver Parameters: For all experiments, we used Gurobi as the MILP solver with a MIP gap accuracy of 0.01% and $\beta = 0.0001$ in constraint (O6). If the solver execution time reaches 144 h then the solver is halted and the best solution found so far is used. Solution times range from typically real time (less than 200s) for optimized adaptive solutions of Network 1, to over 100 h for fixed-time plans of Network 2; however, once the fixed-time solution is found, it can be deployed indefinitely.

What is the impact of modelling lost time delay? Fig. 4.7(a) shows, for Network 1 without light rail, the average delay per vehicle as a function of demand level under optimized adaptive control. The QTM predicted delay for the controller without lost time is considerably lower, but the policies found with QTM incurring lost time show improved performance under microsimulation and closely match the prediction. Figs. 4.7(b) and 4.7(c) show the microsimulation time-distance plots at demand level II for several links along the arterial of Network 1. The y-axis of these plot shows the distance along the street, and the x-axis shows the evolution over time. Each black trace represents the journey of a vehicle along the street. Traffic signals at fixed distances down the street appear as red horizontal dashed lines, where a solid bar represents that the phase is inactive (i.e., the light is red) and traffic cannot pass; otherwise the phase is active. The time-distance plots capture the queueing behaviour of the traffic as each vehicle decelerates when approaching congested traffic, or red or yellow light ahead. When a vehicle is stationary, its trace becomes horizontal. Green represents QTM traffic flow prediction, where darker shades represent regions of higher density. Fig. 4.7(b) shows the controller with QTM incurring lost time delay, and the prediction closely matches the microsimulation. However, in Fig. 4.7(c) the QTM policy without lost time is unrealizable and the microsimulation quickly diverges.

What is the impact of a common cycle time on the controller? Fig. 4.8 shows, for Network 1, with and without light rail, the average delay per vehicle as a function of demand for each of the four controllers. With no light rail (Fig. 4.8(a)) both the optimized adaptive controllers have very similar average delay, while the fixed controllers have the same average delay. An inspection of the fixed signal plans shows them to be identical. With the introduction of the fast light rail (Fig. 4.8(b)) the controllers without common cycle length constraints

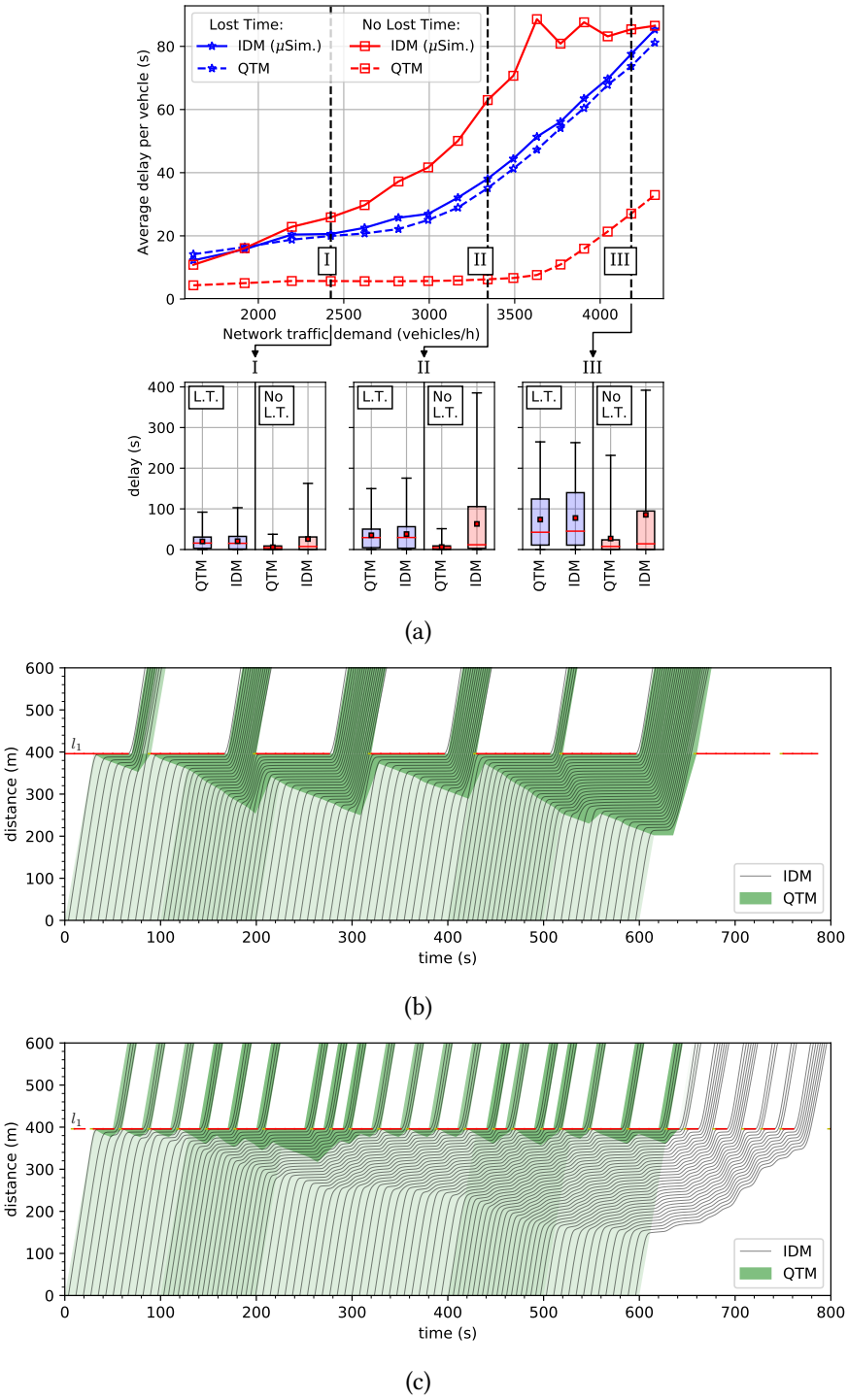


Figure 4.7: Lost time delay. (a) Upper plot shows QTM average delay, with and without lost time, compared to microsimulation of Network 1 (no light rail) at increasing demand levels. Lower plot: Box plots representing distribution of delay at three different demand levels. Policies found with QTM incurring lost time show improved performance under microsimulation. (b,c) Time-distance plots at demand level II, from q_2 to q_4. (b) QTM incurring lost time delay closely predicts microsimulation, but in (c) QTM policies without lost time are unrealizable and the microsimulation quickly diverges.

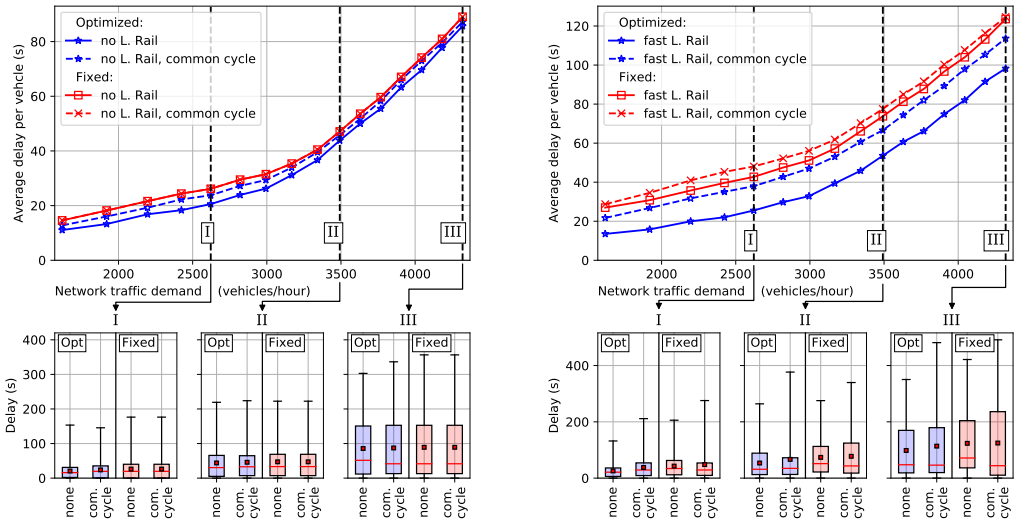


Figure 4.8: Comparison of the four controllers running on Network 1. (a) With no light rail: The optimized controllers are very close in performance, while both the fixed controllers find the same solution (b) With the fast light rail: The controllers without common cycle constraints are able to further reduce delay by using different cycle lengths along the arterial.

are able to reduce delay further by utilizing different cycle lengths along the arterial. The same outcome was observed on Network 2 without light rail, however with the introduction of the light rail, the controllers with common cycle length constraints were unable to find a feasible solution that also satisfied the light rail schedule.

Considering the lower plots that show the distribution of delay and give an indication of the quality of the solutions, we can see that at demand level II in Fig. 4.8(a) and Fig. 4.8(b), both the optimized controllers find policies with similar average delay. But the box plots show that the optimized controller with common cycle length trades a lower median and upper quartile for a higher maximum delay.

These results corroborate with methods already employed by traffic engineers Wolshon and Pande [2016], that for regular traffic networks, using a common cycle lengths between adjacent intersections is a useful aid for achieving good coordination, especially with fixed-time control. However, when the network is not regular, or has addition constraints such as lightrail, improved solutions may be found with mixed cycle lengths.

Is it possible to mitigate the impact of light rail on delay? Figs. 4.9(a) and 4.9(b) show, for each network, the average delay per vehicle as a function of demand for both fixed-time and optimized adaptive control approaches in three scenarios: before the light rail and after the installation of light rail using the slow and the fast schedules. In all cases the controller models lost time. As we hypothesized, optimized adaptive control is able to mitigate the impact of the introduction of light rail and it marginally increases the average delay when compared with the average delay produced by the fixed-time controller **before** the light rail. Moreover, as shown in Figs. 4.9(c) and 4.9(d), the optimized adaptive controller also produces

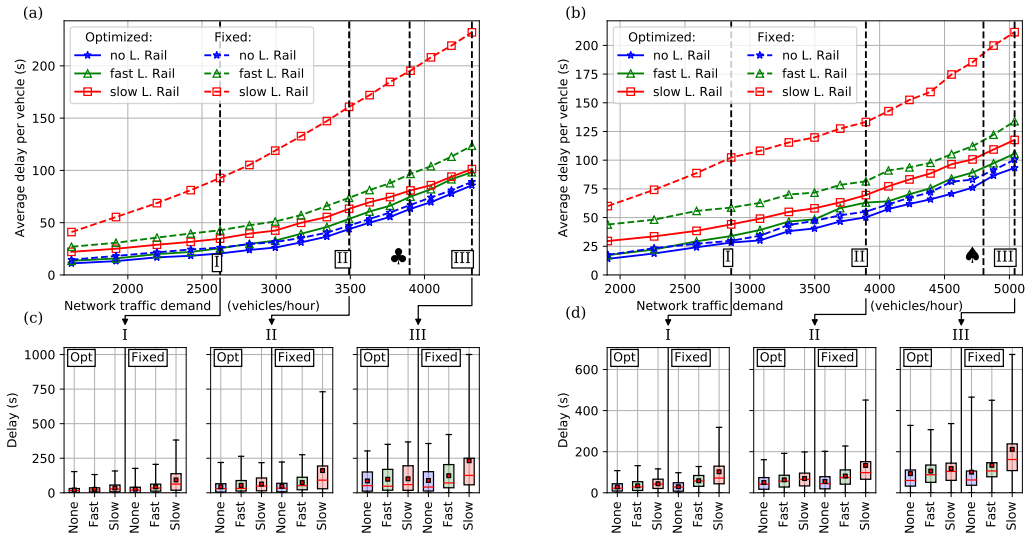


Figure 4.9: Microsimulation average delay by network demand for the Network 1 (a) and Network 2 (b). (c,d): Box plots representing the observed distribution of delay for 3 different values of demand for each network, comparing delay distribution without the light rail with the impact of the fast and slow schedules. Optimized adaptive control is able to mitigate the impact of light rail on average delay, while also producing better signal plans (lower median, third quartile and maximum delay).

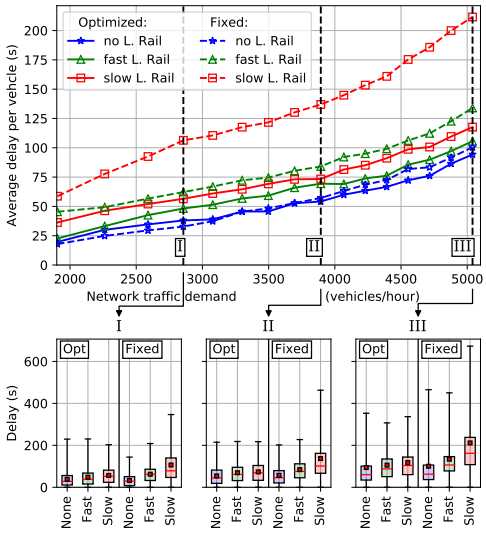


Figure 4.10: Microsimulation average delay for Network 2, averaged over 10 different demand profiles, where the arrival rates differ from the profile used by each controller. Comparing these results (unexpected arrivals) against Figs. 4.9(b) and 4.9(d) (arrivals as expected), we can see that the obtained policies are robust.

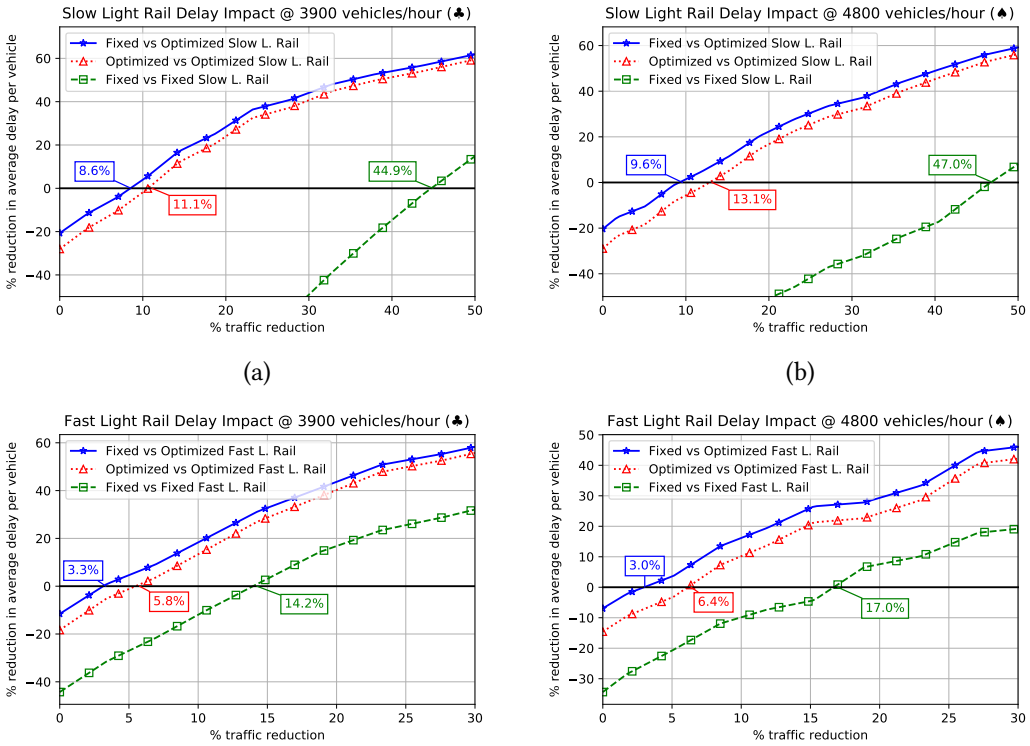


Figure 4.11: Impact on average delay for the Network 1 (first column) and Network 2 (second column) for both light rail schedules (rows) in different scenarios (curves) of traffic control system before and after installation of light rail. The x-axis is the percentage of vehicles switching to the public transportation and the y-axis is the % reduction in delay after the light rail is installed. Negative % represents an increase in average delay. The vehicle demand for (a-d) are marked as ♣ and ♠ in their respective plots in Fig. 4.9.

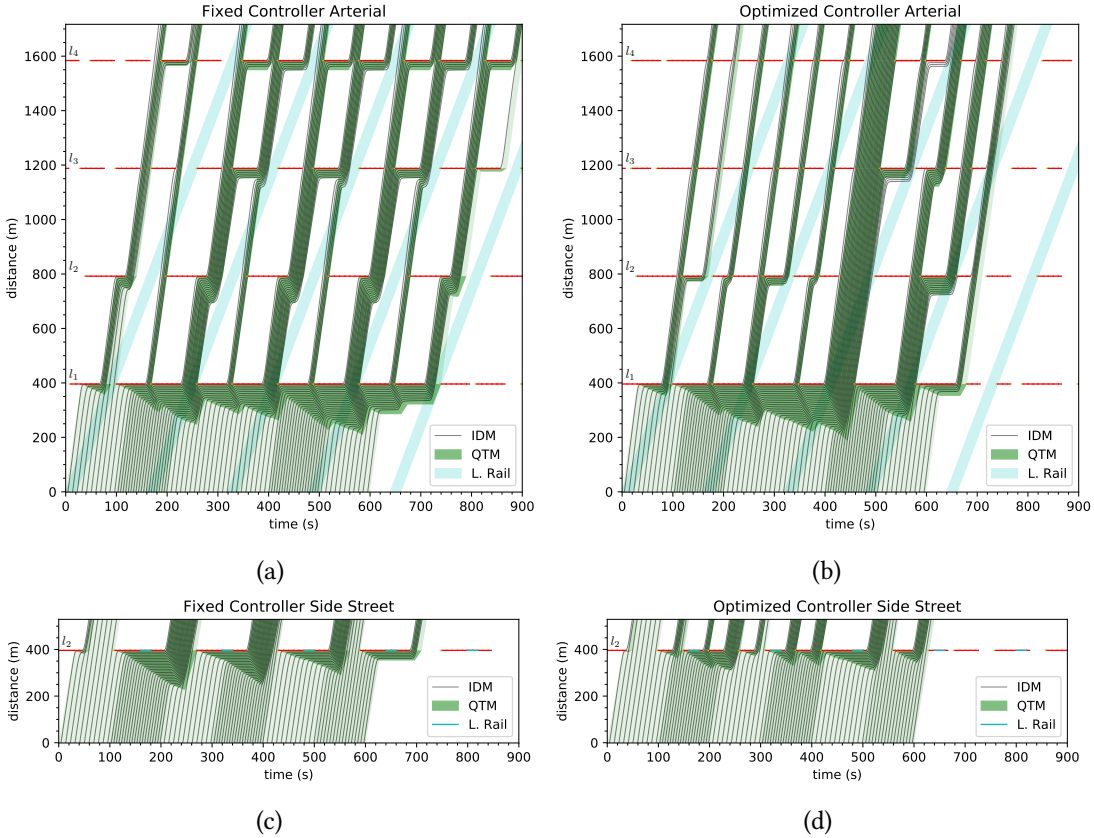


Figure 4.12: (a-b) Micro-simulation time-distance plots from Network 1 with the fast light rail schedule, along the links from q_2 to q_{13}. Both controllers find well coordinated solutions where the timing of green signals along the link is offset at each intersection to maintain a continuous flow of traffic at the free flow speed, a solution well known to traffic engineers. However, the optimized controller is able to dynamically adjust the "width" of the bands to match the traffic volume along the link, allowing more green time to be allocated to cross traffic. (c-d): Microsimulation time-distance plots of side street q_5 to q_6. While both controllers can find coordinated policies along the arterial, the optimized adaptive controller is able to clear out the queues (horizontal flow lines) in the side streets following the transit of the light rail.

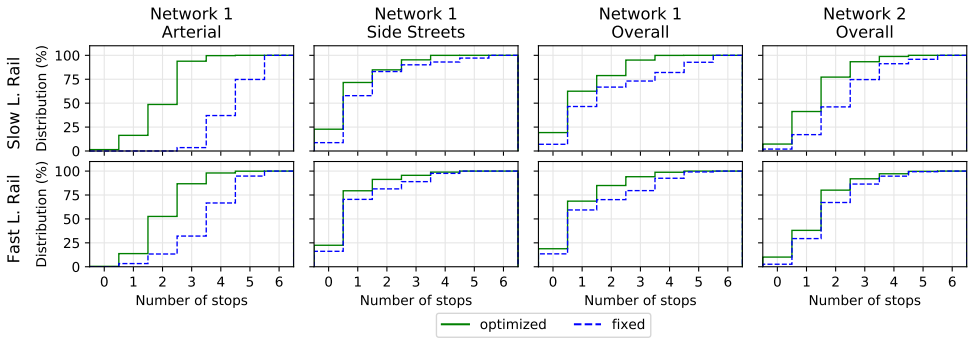


Figure 4.13: Impact of controller on number of stops as a cumulative distribution. Top row is for the slow light rail schedule. Bottom row is for the fast light rail schedule. In all cases the optimized controller does better with less stops at higher frequencies.

	Light Rail	Fixed	Opt. Adapt.	**Improv.**
Arterial (♣) @ 3900 Veh/h	None	66.8 s	62.9 s	**5.7%**
	Fast	96.3 s	74.5 s	**22.7%**
	Slow	194.2 s	80.6 s	**58.7%**
Grid (♠) @ 4800 Veh/h	None	87.3 s	81.5 s	**6.7%**
	Fast	117.4 s	93.4 s	**20.4%**
	Slow	192.9 s	105.1 s	**45.5%**

Table 4.1: Average delay computed via microsimulation in seconds and improvement of optimized adaptive controller over fixed-time controller for both networks and the three light rail scenarios. The demand level is fixed and correspond to the points ♣ and ♠ in Figs. 4.9(a) and 4.9(b), respectively. The improvement obtained by our optimized adaptive approach when a light rail is introduced, ranges from 20.4% to 57.8% w.r.t. the fixed-time approach.

better signal plans than the fixed-time controller, i.e., plans with smaller median, third quartile, and maximum delay.

Table 4.1 shows the average delay in seconds of the optimized adaptive and fixed-time controllers for both networks. In the scenarios with a light rail, the improvement obtained by the optimized adaptive approach ranges from 20.4% to 58.7% w.r.t. the fixed-time approach. We can see that the optimized adaptive controller successfully nullifies the impact of adding a light rail to the networks since the average delay obtained by it is approximately the same as the average delay for the fixed-time controller with no light rail, with the average delay increased by at most 17.8 s.

What is the impact of unexpected arrivals on the controller? Fig. 4.10 shows the average delay per vehicle averaged over 10 demand profiles for Network 2, where the arrival rates in the simulation differ over time from the profile used to generate the signal plan. The overall conclusion is still the same: the optimized controller outperforms the fixed-time controller. Compared to the results in Fig. 4.9(b), the optimized controller shows a slight increase in the average delay at lower demand levels and almost no change at higher demand levels while, for the fixed controller, the policy is more robust to the unexpected arrivals

and there is minimal change across all demand levels. The fixed controller is more robust to unexpected arrivals because its policy is optimize for the average arrival rate as opposed to an exact demand profile, thus any demand profile with the same or similar average arrival rate will have little impact in the average delay. Since the optimized controller attempts to coordinate signal timings with expected changes in inflow rate to form platoons, the changes in the simulated demand profile have an impact the average delay. The impact of unexpected arrivals at lower demand levels is small because some vehicles may need to wait an additional cycle before joining a platoon coordinated by the policy, while at higher demand levels, the residual queue build up at the inputs buffers any mismatch resulting in a negligible impact in the average delay. In all cases, the box plots show the solution quality for each controller is not impacted by the unexpected arrivals.

How many drivers must switch to using light rail to maintain the same average delay? This is a question that will be asked by planners evaluating the impact of adding a light-rail to a traffic network, along with an upgrade to the signal control system. Fig. 4.11 shows the percentage reduction in average delay as a function of the percentage of vehicles who's drivers are switching to traveling on the light rail. In these plots, the demand level is fixed and higher values are better (i.e., there is a larger decrease in the average delay) and zero means that there is no change after installing light rail. For the three combinations of before and after policies presented, we can see that, while keeping the fixed-time controller requires from 14.2% to 47% of the drivers to switch to light rail in order to obtain the same average delay as before its installation, the optimized adaptive approach requires only from 5.8% to 13.1% of the drivers to switch when already using optimized adaptive control **before** the light rail. When compared to fixed-time before the light rail and optimized adaptive after, the gains are even greater with only 3% to 9.6% of the drivers required to switch to the light rail.

How does the quality of optimized adaptive policies compare with fixed policies? To answer this we show in Figs. 4.12(a) to 4.12(d) the microsimulation time-distance plots for several streets in Network 1. Figs. 4.12(a) and 4.12(b) show that both controllers balance between establishing coordinated "green corridors" along the arterial, and servicing the side streets, where the combined density at times exceeds that of the arterial. However, the fixed controller is forced to find a single repeating policy sized for the average traffic density in the network. As a byproduct, the side street (Fig. 4.12(c)) under the fixed-time controller suffers from accumulative queue build-up following each transit of the light rail. In Figs. 4.12(b) and 4.12(d), we see that the optimized adaptive controller is able to clear out the queue build up in the side street by increasing the phase time of the side street for a cycle after the transit has passed through, and then returns to a schedule that prioritizes the arterial depending on the changes in traffic density. When the traffic density in the arterial is higher than the side streets, the optimized adaptive controller will coordinate "green corridors" along the arterial.

Fig. 4.13 provides more details on the behavior of the signal plans for demand level II (Figs. 4.9(a) and 4.9(b)) by showing the cumulative number of vehicles by number of observed stops. In all cases for Network 1 and 2 the optimized controller does better with less stops at higher frequencies. For Network 1, we see that both controllers choose to prioritize the side streets over the arterial, with less stops at higher frequencies in the side streets. But in the case of the slow light rail with the optimized controller, 94% of the vehicles experi-

ence three or less stops along the arterial while for the fixed controller 100% of the vehicles experience three stops or more.

4.4 Conclusion

In this chapter, we introduced a new method to generate an adaptive controller that optimizes traffic signals integrated with light rail schedule constraints. The obtained adaptive controllers are guaranteed to be globally optimal and, to the best of our knowledge, this is the first globally optimal algorithm capable of handling light rail schedule constraints. Our approach is based on the Queue Transmission Model of traffic signal control which we extended to incorporate light rail schedule constraints. We also provided a novel way to model lost time directly as a signal timing constraint and show that it is critical to finding optimized signal plans.

We also introduced a novel approach to compute fixed-time controller plans that optimize cycle times, phase splits and offsets. The obtained fixed-time controllers are also guaranteed to be globally optimal and they can handle both light rail schedule constraints and common cycle length constraints. The computed fixed-time control schedules can be incorporated *immediately* into existing fixed-time traffic controller infrastructure, yielding important benefits for those municipalities that prefer not to migrate to a fully adaptive control.

Lastly, we have compared our optimal adaptive and fixed-time controllers in a comprehensive suit of experiments using microsimulation as a realistic, finer-grained, nonlinear model of traffic flow. Our results show that the optimal adaptive controller is able to minimize the impact of introducing light rail on conventional traffic networks on the average delay with respect to fixed-time signal control. The experiments also show that the adaptive controllers finds better quality solutions, i.e., solutions with substantially lower third quartile and maximum observed delay. Our key results demonstrate for the first time the potential of MILP-based QTM traffic signal control approaches to virtually nullify the impact of installing light rail on conventional traffic — our model can reduce traffic delay by up to 58.7% over optimal fixed-time control when light rail is introduced. Consequently, the use of MILP-based optimized adaptive controllers like QTM could remove the critical public concern of increased traffic delay resulting from light rail installation, and thus positively impact the environment, urban productivity, and commute time reductions for all commuters.

For future work, a key question to resolve is how large we can scale the traffic and light rail network before we need to investigate decomposition-based approaches to scaling the solution (e.g., MILP-based methods like dual decomposition or region-based traffic network partitioning schemes). Future work should also examine the (online) learnability of QTM parameters from different traffic sensor data, for instance, conventional inductive (double) loop counters, radar, and video feeds. Finally, noting that the nonlinear microsimulation model offers a higher-fidelity model of traffic behavior, future work should consider expanding the QTM to model nonlinear traffic flows Lu et al. [2011]; Muralidharan et al. [2009]; Kim [2002]; Huang [2011] and investigating the benefits of nonlinear optimization relative to the existing QTM.

Chapter 5

QTM Equivalence

5.1 Overview

The models CTM, LTM and VT presented in Chapter 2 are all derived from LWR theory, while QTM presented in Chapter 3 is based on a simple point queue model,. Yet, the micro-simulation results in Chapter 4 suggest that QTM is modelling the fundamental dynamics of traffic flow along links and through signalised intersections, with reasonable fidelity. In this chapter we will investigate why, and use the surprising results to develop further extensions of QTM.

First we establish a theoretical framework to show equivalences between the models and their objective functions. We then use this result to show that the problem of vehicle withholding is an artefact of the solution method and has no impact on solution optimality. We then extend QTM with higher fidelity modeling of LWR traffic flows, and in turn, use QTM to extend VT with turning traffic and show that it is a stochastic shortest path problem, equivalent to finding the optimal value function of a Markov Decision Problem. Finally, we present a novel reformulation of QTM as a continuous time solver that can find *exact* solutions to LWR theory.

5.2 Objective Function Equivalence

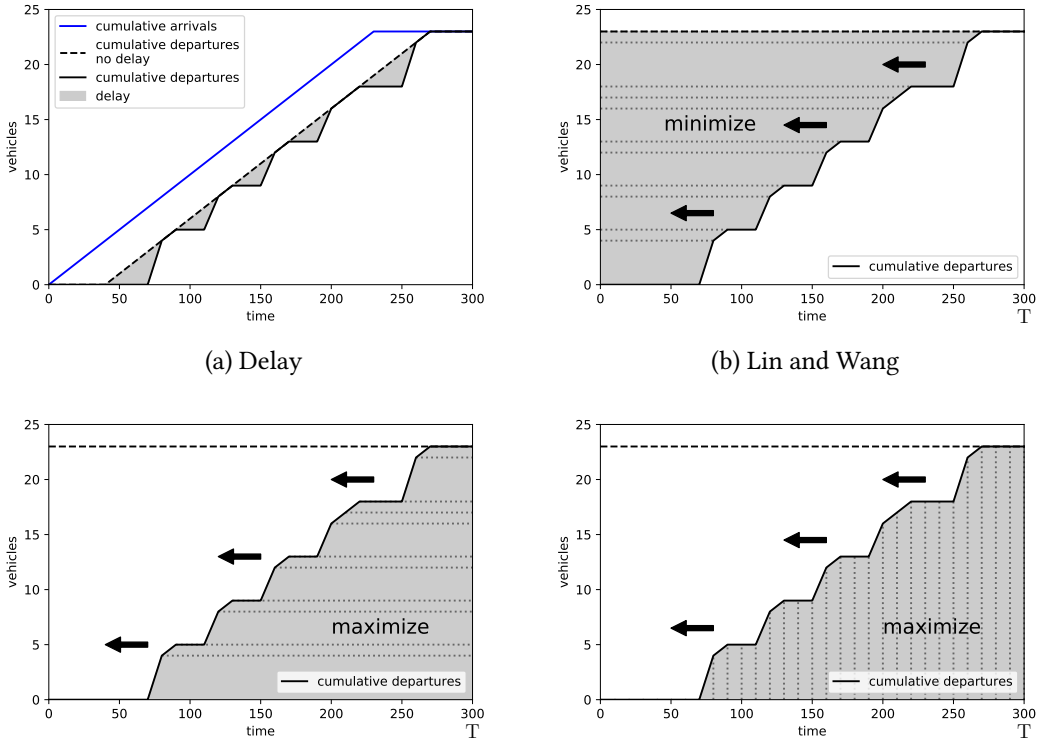
Vehicle delay in a traffic link is the area between the cumulative departure curve for the link and the cumulative arrival curve projected forward by the link free flow travel time (Fig. 5.1(a)). The MILP formulations of QTM, CTM [Lin and Wang, 2004], and VT [Wada et al., 2017], seek to minimize this delay through the objective function. Lin and Wang [Lin and Wang, 2004] minimize the total travel time of all vehicles in the network with the term (2.33). Since $y^{\text{out}}_{i,t}$, the number of vehicles leaving destination cell i during interval t, is equivalent to $q^{\text{out}}_{i,n} = \Delta t_n f^{\text{out}}_{i,n}$, the number of vehicles departing the network from i during interval n, we can formulate (2.33) in QTM terms with (5.1).

$$\min \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} t_n q^{\text{out}}_{i,n} \tag{5.1}$$

The objective (5.1) minimizes the area above the cumulative departure curves (Fig. 5.1(b)). We can convert this into a maximization by noting that the area below the cumulative departure curve is equal to the total rectangular area enveloping the curve, less the area above, which is formulated with (5.2).

$$\sum_{n=1}^{N} \mathrm{T} q^{\text{out}}_{i,n} - \sum_{n=1}^{\mathrm{N}} t_n q^{\text{out}}_{i,n} = \sum_{n=1}^{\mathrm{N}} (\mathrm{T} - t_n)\, q^{\text{out}}_{i,n} \tag{5.2}$$

Thus by maximizing the area *below* the departure curve (Fig. 5.1(c)), we minimize the area above and the maximization objective (5.3) of QTM is equivalent to the minimization objective



(a) Delay

(b) Lin and Wang

(c) QTM

(d) Wada *et al.*

Figure 5.1: Cumulative curve objective functions. (a) Vehicle delay is the area between the cumulative departure curve and the cumulative arrival curve projected forward by the travel time. (b) Lin and Wang minimize delay by minimizing the sum of the horizontal rectangles above the departure curve. (c) QTM minimizes delay by maximizing the sum of the horizontal rectangles below the departure curve. (d) Wada *et al.* minimize delay by maximizing the sum of the vertical rectangles below the departure curve.

of Lin and Wang (5.1) and (2.33).

$$\max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} (\mathrm{T} - t_n)\, \Delta t_n f_{i,n}^{\text{out}} \tag{5.3}$$

[Wada et al., 2017] also minimize delay by maximizing the area under the cumulative departure curve (Fig. 5.1(d)) using the objective function (5.4), where $N_{i,n}$ is the value of the departure curve for i at the end of interval n.

$$\max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} \Delta t_n N_{i,n}^{\text{out}} \tag{5.4}$$

This is equivalent to the QTM objective (5.3), and, given $t_n = \sum_{k=1}^{n} \Delta t_k$, and the that value of the departure curve for i at the end of interval n is given by $N_{i,n}^{\text{out}} = \sum_{k=1}^{n} q_{i,k}^{\text{out}}$. (5.4) can be derived directly from (5.3) using equation (5.5).

$$\begin{aligned}
\max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} (\mathrm{T} - t_n)\, \Delta t_n f_{i,n}^{\text{out}} &= \max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} \left(\mathrm{T} - \sum_{k=1}^{n} \Delta t_k \right) \Delta t_n f_{i,n}^{\text{out}} \\
&= \max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} \left(\sum_{k=n}^{\mathrm{N}} \Delta t_k \right) \Delta t_n f_{i,n}^{\text{out}} \\
&= \max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} \sum_{k=1}^{n} \Delta t_n \left(\Delta t_k f_{i,k}^{\text{out}} \right) \\
&\qquad \left(\text{since } \sum_{n=1}^{\mathrm{N}} \sum_{k=n}^{\mathrm{N}} a(n,k) = \sum_{n=1}^{\mathrm{N}} \sum_{k=1}^{n} a(k,n) \right) \\
&= \max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} \Delta t_n N_{i,n}^{\text{out}}
\end{aligned} \tag{5.5}$$

Thus, we see that the QTM, CTM and VT formulations all equivalently minimize the total travel time in the network through their respective objective functions.

5.3 VT Equivalence

In Section 2.7 a method to solve the LWR equation using Variational Theory was presented. When evaluated with discrete time steps, VT takes the form of a constrained shortest path problem on the boundary conditions of the traffic network, which can be solved using the MILP (O4, V2). This result suggests a framework for demonstrating whether a given traffic MILP finds solutions to LWR theory, by constructing an equivalence to the VT constraints (O4, V2).

5.3.1 QTM Equivalence

If we restrict ourselves to traffic flows without turns and homogeneous time steps, then an equivalence can be derived between QTM and VT. To show that QTM has an embedded VT constrained shortest path problem, we need to connect QTM constraints to nodes on a VT

network with forward, backward and horizontal wave paths (an extension for multiple waves is given in Section 5.5). If x_i^{in} and x_i^{out} are the distances to the input and exit of queue i, and $t_n = \sum_{k=1}^{n-1} \Delta t_k$, then MILP (5.6) – (5.9) represents a VT constrained shortest path problem, expressed in QTM constants $\mathrm{T}_i^{\mathrm{p}}$, Q_i and $\mathrm{F}_{i,j}$.

$$\max \sum_{n=1}^{\mathrm{N}} \sum_{i=1}^{|\mathcal{Q}|} N(x_i^{\text{out}}, t_n) \tag{5.6}$$

$$\text{subject to:} \quad N(x_i^{\text{out}}, t_n) \leq N(x_i^{\text{in}}, t_n - \mathrm{T}_i^{\mathrm{p}}) \tag{5.7}$$

$$N(x_i^{\text{in}}, t_{n-1}) \leq N(x_i^{\text{out}}, t_n) + \mathrm{Q}_i \tag{5.8}$$

$$N(x_i^{\text{out}}, t_n) \leq N(x_i^{\text{out}}, t_{n-1}) + \Delta t_{n-1} \mathrm{F}_{i,j} p_{i,j,n-1} \tag{5.9}$$

Without loss of generality we assume for all i, $q_{i,1} = 0$, such that there is an epoch at t_1 where the network is free of traffic. Since the original Variational Theory described by Deganzo in [Daganzo, 2005a] does not include turning traffic, we will restrict the derivation to traffic networks where all the outflow from a queue i flows into queue j, giving $q_{i,n}^{\text{out}} = q_{j,n}^{\text{in}} = \Delta t_n f_{i,j,n}$ from constraints (Q3) and (Q4). For simplicity, and similar to [Wada et al., 2017], we will also assume that there is a single binary phase, $p_{i,j,n}$, controlling each queue. To ensure that the VT network is fully connected, Δt_n must be homogeneous and there must be a positive integer $m = {}^{\mathrm{T}_i^{\mathrm{p}}}/_{\Delta t_n} \in \mathbb{Z}^+$, such that $t_{n-m} = t_n - \mathrm{T}_i^{\mathrm{p}}$. Then we can rewrite QTM constraints (Q5) and (Q6) in terms of m as (5.10) and (5.11), and the cumulative vehicle counts at the exit of queue i at time t_n, and input of queue i at time t_{n+1} and $t_n - \mathrm{T}_i^{\mathrm{p}}$ as (5.12) – (5.14).

$$q_{i,n} = q_{i,n-1} - q_{i,n-1}^{\text{out}} + q_{i,n-m-1}^{\text{in}} \tag{5.10}$$

$$q_{i,n} \leq \mathrm{Q}_i - \sum_{k=n-m}^{n} q_{i,k}^{\text{in}} \tag{5.11}$$

$$N(x_i^{\text{out}}, t_n) = \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \tag{5.12}$$

$$N(x_i^{\text{in}}, t_{n+1}) = \sum_{k=1}^{n} q_{i,k}^{\text{in}} \tag{5.13}$$

$$N(x_i^{\text{in}}, t_n - \mathrm{T}_i^{\mathrm{p}}) = \sum_{k=1}^{n-m-1} q_{i,k}^{\text{in}} \tag{5.14}$$

To construct a forward wave path constraint, we first rearrange (5.10) and take the cumulative sum of both sides from t_1 to get (5.15). Then expanding and simplifying the left hand

side of (5.15) gives (5.17).

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} = \sum_{k=1}^{n-1} \left(q_{i,k} - q_{i,k+1} + q_{i,k-m}^{\text{in}} \right) \tag{5.15}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} = q_{i,1} - q_{i,n} + \sum_{k=1}^{n-m-1} q_{i,k}^{\text{in}} \tag{5.16}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} = -q_{i,n} + \sum_{k=1}^{n-m-1} q_{i,k}^{\text{in}} \qquad (\text{since } q_{i,1} = 0) \tag{5.17}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \leq \sum_{k=1}^{n-m-1} q_{i,k}^{\text{in}} \qquad (\text{since } q_{i,n} \geq 0) \tag{5.18}$$

Since $q_{i,n} \geq 0$ we can relax the equality in (5.17) to get (5.18) and it is easy to see that this is the forward wave path constraint (5.7), which replaces the QTM constraints (Q2) – (Q5).

For the horizontal wave paths we start with the cumulative sum over $q_{i,n}^{\text{out}}$ in (5.19). Since $q_{i,n-1}^{\text{out}} \leq \Delta t_{n-1} \mathrm{F}_{i,j} p_{i,j,n-1}$, we can then relax the equality in (5.19) to get (5.20), which is the horizon wave path constraint (5.9).

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} = q_{i,n-1}^{\text{out}} + \sum_{k=1}^{n-2} q_{i,k}^{\text{out}} \tag{5.19}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \leq \Delta t_{n-1} \mathrm{F}_{i,j} p_{i,j,n-1} + \sum_{k=1}^{n-2} q_{i,k}^{\text{out}} \qquad (\text{since } q_{i,n-1}^{\text{out}} \leq \Delta t_{n-1} \mathrm{F}_{i,j} p_{i,j,n-1}) \tag{5.20}$$

$$\tag{5.21}$$

To construct the backwards wave path, we rearrange (5.10) and take the cumulative sum of both sides to get (5.22). Expanding and simplifying gives (5.24), and substituting in constraint (Q6), we get (5.25). Finally we can see that (5.26) is the backwards wave path constraint (5.8).

$$\sum_{k=1}^{n-1} q_{i,k-m}^{\text{in}} = \sum_{k=1}^{n-1} \left(q_{i,k+1} - q_{i,k} + q_{i,k}^{\text{out}} \right) \tag{5.22}$$

$$\sum_{k=1}^{n-m-1} q_{i,k}^{\text{in}} = q_{i,n} - q_{i,1} + \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \tag{5.23}$$

$$\sum_{k=1}^{n-m-1} q_{i,k}^{\text{in}} = q_{i,n} + \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \qquad (\text{since } q_{i,1} = 0) \tag{5.24}$$

$$\sum_{k=1}^{n-m-1} q_{i,k}^{\text{in}} \leq \mathrm{Q}_i - \sum_{k=n-m}^{n} q_{i,k}^{\text{in}} + \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \qquad \left(\text{since } q_{i,n} \leq \mathrm{Q}_i - \sum_{k=n-m}^{n} q_{i,k}^{\text{in}}\right) \tag{5.25}$$

$$\sum_{k=1}^{n} q_{i,k}^{\text{in}} \leq \mathrm{Q}_i + \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \tag{5.26}$$

Finally, to show that the QTM objective function is equivalent, we need to derive (5.6) from (O6). From QTM constraint (Q4) we can see that $q_{i,n}^{\text{out}} = f_{i,n}^{\text{out}} + f_{i,j,n}$, and that the left hand side

of (5.27) is equivalent to (O6) with $\beta = 1$.

$$\max \sum_{n=1}^{\mathrm{N}} \sum_{i=1}^{|\mathcal{Q}|} (\mathrm{T} - t_n)\, q_{i,n}^{\text{out}} = \max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} \left(\mathrm{T} - \sum_{k=1}^{n-1} \Delta t_k \right) q_{i,n}^{\text{out}} \tag{5.27}$$

$$= \max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} \left(\sum_{k=n}^{\mathrm{N}-1} \Delta t_k \right) q_{i,n}^{\text{out}} \qquad \left(\text{since } \mathrm{T} = \sum_{k=1}^{\mathrm{N}-1} \Delta t_k\right)$$

$$= \max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} \sum_{k=1}^{n-1} \Delta t_n q_{i,k}^{\text{out}} \qquad \left(\text{since } \sum_{n=1}^{\mathrm{N}} \sum_{k=n}^{\mathrm{N}} a(n,k) = \sum_{n=1}^{\mathrm{N}} \sum_{k=1}^{n} a(k,n)\right) \tag{5.28}$$

Since Δt_n is homogeneous and constant, it can be dropped from (5.28) and we can see that (5.28) is equivalent to (5.6). Then we can use the QTM constraints (O6, Q1–Q6) and the relationships $q_{i,n} \geq 0$, $q_{i,n} \leq \mathrm{Q}_i$, and $q_{i,n}^{\text{out}} \leq \Delta t \mathrm{F}_i p_{\ell,n}$, to derive constraints (5.30) to (5.32), where m is the index such that $t_{n-m} = t_n - \mathrm{T}_i^{\mathrm{p}}$.

$$\max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} (\mathrm{T} - t_n)\, \Delta t_n f_{i,n}^{\text{out}} \tag{5.29}$$

$$\text{subject to:} \qquad \sum_{k=1}^{n} q_{i,k}^{\text{out}} \leq \sum_{k=1}^{n-m} q_{i,k}^{\text{in}} \tag{5.30}$$

$$\sum_{k=1}^{n} q_{i,k}^{\text{out}} \leq \Delta t \mathrm{F}_i p_{\ell,n} + \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \tag{5.31}$$

$$\sum_{k=1}^{n} q_{i,k}^{\text{in}} \leq \mathrm{Q}_i + \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \tag{5.32}$$

The MILP given by constraints (5.29) to (5.32) demonstrates that QTM with homogeneous time steps and no turning traffic is solving a constrained shortest path problem. This result also shows that QTM is modeling an LWR traffic flow with a trapezoidal fundamental diagram, where the forward wave speed is given by $v = \frac{L_i}{\mathrm{T}_i^{\mathrm{p}}}$ and the backwards wave speed is given by $w = \frac{L_i}{\Delta t_n}$, where L_i is the length of link i.

5.3.2 CTM Equivalence

Using a similar method, we can derive a VT equivalence for CTM. First we need to connect CTM constraints to nodes on a VT network connected with forward and horizontal and backwards wave paths. If x_i and x_i^{out} are the distances to the input and exit of cell i, then (5.33) –

(5.34) represents a VT shortest path problem.

$$\max \sum_{n} \sum_{i \in \text{exit}} N(x_i^{\text{out}}, t_n) \tag{5.33}$$

$$\text{subject to:} \quad N(x_{i+1}, t+1) \leq N(x_i, t) \tag{5.34}$$

$$N(x_i, t+1) \leq N(x_i, t) + \text{Q} \tag{5.35}$$

$$N(x_i, t+1) \leq N(x_{i+1}, t) + \text{N} \tag{5.36}$$

With out loss of generality, we assume for the initial conditions that $N(x_i, 0) = 0$, and $n_{i,0} = 0$, which is to say that there is an epoch where all the cells are empty. Then consider the following:

$$N(x_i, t) = \sum_{k=0}^{t-1} y_{i,k} \tag{5.37}$$

$$N(x_i^{\text{out}}, t) = \sum_{k=0}^{t-1} y_{i,k}^{\text{out}} \tag{5.38}$$

$$n_{i,t+1} = n_{i,t} + y_{i,t} - y_{i+1,t} \tag{5.39}$$

$$y_{i+1,t} = n_{i,t} - n_{i,t+1} + y_{i,t} \tag{5.40}$$

$$y_{i,t} = y_{i+1,t} - n_{i,t} + n_{i,t+1} \tag{5.41}$$

$$y_{i,t} = \min \left\{ n_{i-1,t},\ \text{Q},\ \frac{w}{v} \left(\text{N} - n_{i,t} \right) \right\} \tag{5.42}$$

From (5.42) we know that,

$$y_{i+1,t} \leq n_{i,t} \tag{5.43}$$

since all these vehicles must flow through the cell to be counted. For the forward wave we have,

$$\sum_{k=0}^{t-1} y_{i+1,k} = \sum_{k=0}^{t-1} \left(n_{i,k} - n_{i,k+1} + y_{i,k} \right)$$

$$\sum_{k=0}^{t-1} y_{i+1,k} = n_{i,0} - n_{i,t} + \sum_{k=0}^{t-1} y_{i,k}$$

$$n_{i,t} + \sum_{k=0}^{t-1} y_{i+1,k} = n_{i,0} + \sum_{k=1}^{t-1} y_{i,k}$$

$$y_{i+1,t} + \sum_{k=0}^{t-1} y_{i+1,k} \leq \sum_{k=0}^{t-1} y_{i,k} \qquad \text{(since } y_{i+1,t} \leq n_{i,t} \text{ and } n_{i,0} = 0\text{)}$$

$$\sum_{k=0}^{t} y_{i+1,k} \leq \sum_{k=0}^{t-1} y_{i,k} \tag{5.44}$$

$$\implies N(x_{i+1}, t+1) \leq N(x_i, t)$$

And for the signal wave, we have:

$$\sum_{k=0}^{t} y_{i,k} = y_{i,t} + \sum_{k=0}^{t-1} y_{i,k}$$

$$\sum_{k=0}^{t} y_{i,k} \leq \mathrm{Q} + \sum_{k=0}^{t-1} y_{i,k} \qquad \text{(since } y_{i,t} \leq \mathrm{Q}\text{)}$$

$$\implies N(x_i, t+1) \leq N(x_i, t) + \mathrm{Q} \tag{5.45}$$

We can derive the backwards wave constraint for the case $\frac{w}{v} = 1$. Equivalence for backwards wave does not hold for the case $\frac{w}{v} \neq 1$.

$$\sum_{k=0}^{t-1} y_{i,k} = \sum_{k=0}^{t-1} \left(y_{i+1,k} - n_{i,k} + n_{i,k+1} \right)$$

$$\sum_{k=0}^{t-1} y_{i,k} = n_{i,t} - n_{i,0} + \sum_{k=0}^{t-1} y_{i+1,k}$$

$$y_{i,t} + \sum_{k=0}^{t-1} y_{i,k} = y_{i,t} + n_{i,t} + \sum_{k=0}^{t-1} y_{i+1,k}$$

$$\sum_{k=0}^{t} y_{i,k} \leq \frac{w}{v} \left(\mathrm{N} - n_{i,t} \right) + n_{i,t} + \sum_{k=0}^{t-1} y_{i+1,k} \qquad \text{(since } y_{i,t} \leq \tfrac{w}{v} \left(\mathrm{N} - n_{i,t} \right)\text{)}$$

$$\sum_{k=0}^{t} y_{i,k} \leq \mathrm{N} + \sum_{k=0}^{t-1} y_{i+1,k} \qquad \text{(since } n_{i,t} \leq \mathrm{N}\text{)}$$

$$\implies N(x_i, t+1) \leq N(x_{i+1}, t) + \mathrm{N} \tag{5.46}$$

We can then derive an equivalent constrained shortest path MILP using CTM constraints (O2, C1–C4, C8, C10, C12, C24, C26), as the MILP given by constraints (5.47) to (5.50).

$$\max \sum_{i \in \text{exit}} \sum_{t} (\mathrm{T} - t)\, y^{\text{out}}_{i,t} \tag{5.47}$$

$$\text{subject to:} \quad \sum_{k=0}^{t} y_{i+1,k} \leq \sum_{k=0}^{t-1} y_{i,k} \tag{5.48}$$

$$\sum_{k=0}^{t} y_{i,k} \leq p_{\ell,t} \mathrm{Q} + \sum_{k=0}^{t-1} y_{i,k} \tag{5.49}$$

$$\sum_{k=0}^{t} y_{i,k} \leq \mathrm{N} + \sum_{k=0}^{t-1} y_{i+1,k} \tag{5.50}$$

5.3.3 LTM Equivalence

To derive an equivalence for LTM, first we restrict the traffic flows to be without turns. In which case we have all the sending flow from link j received by link i and $R_{i,n} = S_{j,n}$. Therefore $N^{\text{in}}_{i,n} = N^{\text{out}}_{j,n} = N(x_i, t)$, and we get $S_{i,n} = N^{\text{out}}_{i,n} - N^{\text{out}}_{i,n-1}$ and $R_{i,n} = N^{\text{in}}_{i,n} - N^{\text{in}}_{i,n-1}$ from constraints (L9) to (L14). Then by substituting and rearranging, we get the forward wave path constraint (5.52) from (L6), the signal wave path constraint (5.53) from (L8), and the

backwards wave path constraint (5.54) from (L2).

$$\max \sum_{i \in \text{exit}} \sum_{n} N^{\text{out}}_{i,n} \tag{5.51}$$

$$\text{subject to:} \quad N^{\text{out}}_{i,n} \leq N^{\text{in}}_{i,n-v_i} \tag{5.52}$$

$$N^{\text{out}}_{i,n} \leq N^{\text{out}}_{i,n-1} + p_{i,j,n} \mathrm{F}_{i,j} \tag{5.53}$$

$$N^{\text{in}}_{i,n} \leq N^{\text{out}}_{i,n-w_i} + c_i \tag{5.54}$$

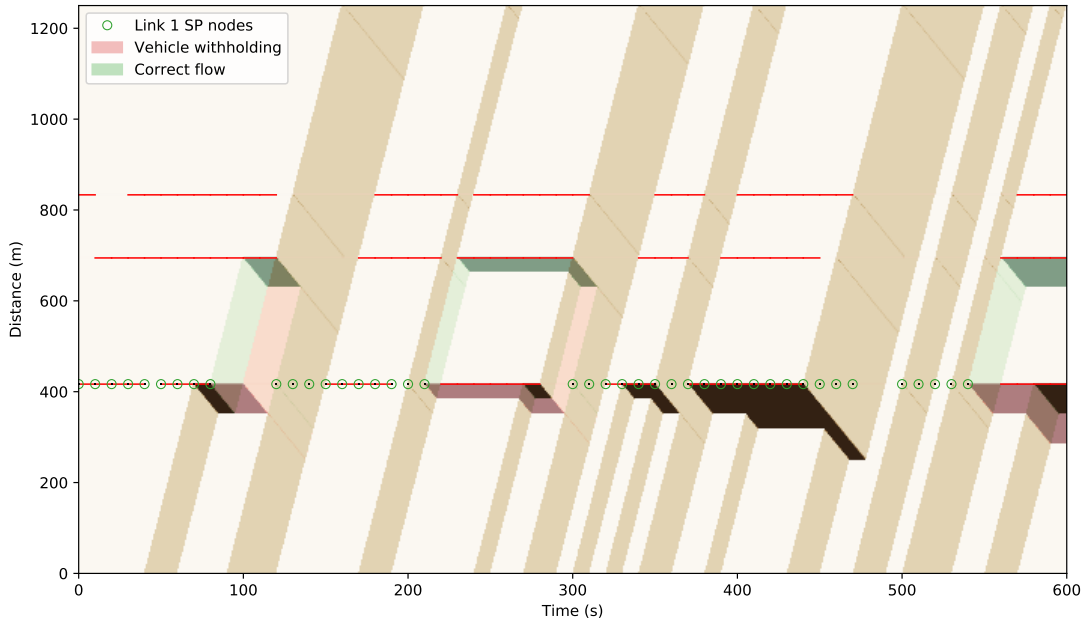
The MILP given by constraints (5.51) to (5.54) solves LTM as a constrained shortest path problem. It is interesting to note that in an LTM solution, the states of $z^{\text{in}}_{i,n}$ and $z^{\text{out}}_{i,n}$ will then indicate the shortest path wave choice at each node. If link i is upstream of link j, $z^{\text{in}}_{j,n} = 0$ for backwards waves, $z^{\text{out}}_{i,n} = 0$ for forward waves, and $z^{\text{out}}_{i,n} = 1$ for signal waves.

5.4 Traffic Withholding

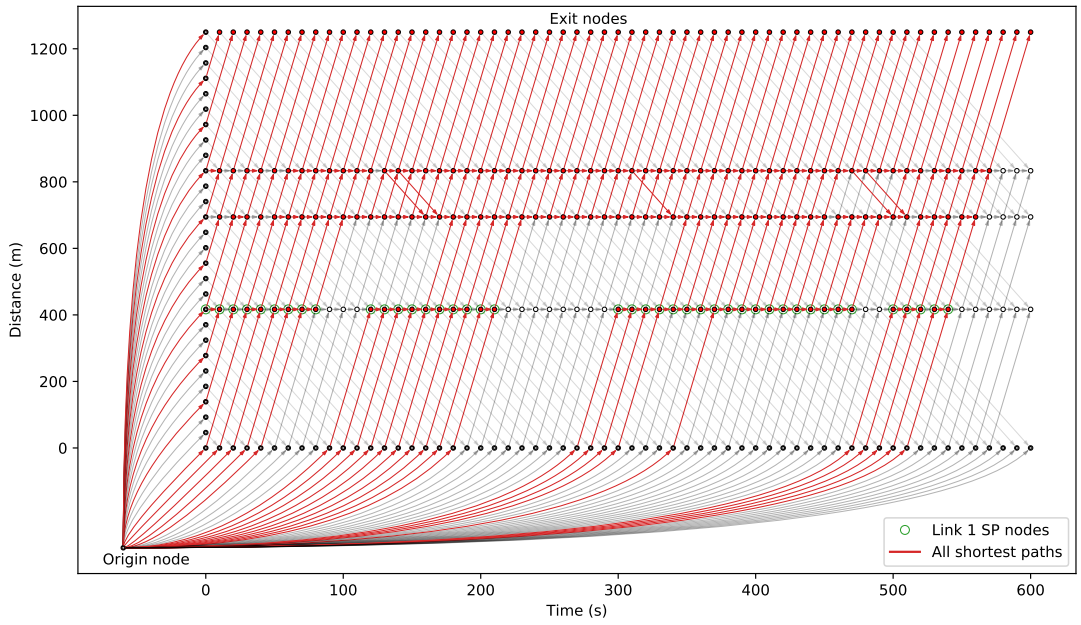
Traffic withholding is a consequence of the relaxed lower bounds in QTM constraints (Q1) and (Q2) and on CTM constraints (C2) to (C4), such that traffic may not move downstream, even if it is clear to do so. To address this, Lin and Wang [Lin and Wang, 2004] add an additional term to the objective function with a small weight to hold the cell outflow against the upper bound. Similarly, QTM also uses a weighted term in the objective function to address any possible withholding of the queue outflow.

However, equations (5.1), (5.2), (5.3) and (5.5) show us that the objective functions of QTM and CTM, without the withholding terms, are maximizing the values of the cumulative vehicle counts at the exit. And we know from the VT equivalences in the previous section, that the QTM and CTM MILPS are also determining the values by finding the shortest path from initial–boundary nodes to each exit nodes. An important observation here is that to optimize the signal plan, the solver needs only to find the correct values of the flows along the shortest paths to the exit nodes of the network. If the withholding term has been removed from the QTM and CTM objective functions (O6) and (O2), then withholding can only occur at those nodes in the equivalent VT lattice that are not on a shortest path to an exit node, and therefore do not impact on the optimality of the solution.

Figure 5.2 gives an example of a traffic flow with vehicle withholding. Figure 5.2(a) shows that the flow between 80 s and 120 s withholds during the green phase at the first traffic signal, and again between 210 s and 220 s, and 540 s and 560 s. The withholding flow is shown in red and the correct flow is shown in green. The VT network and solution for the flow is shown in Fig. 5.2(b), with all possible shortest paths between the origin node and all the exit nodes are enumerated. Those nodes at the first traffic signal on a shortest path are circled green in both Fig. 5.2(a) and Fig. 5.2(b). We can see that the regions in Fig. 5.2(a) where withholding occurs, are restricted to the nodes that are not on any possible shortest path, and therefore do not contribute to the objective value.

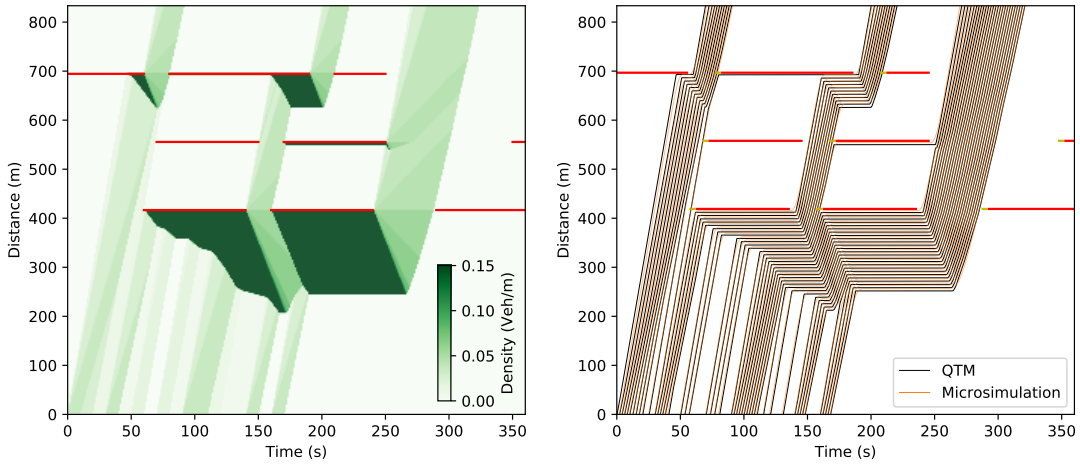


(a)

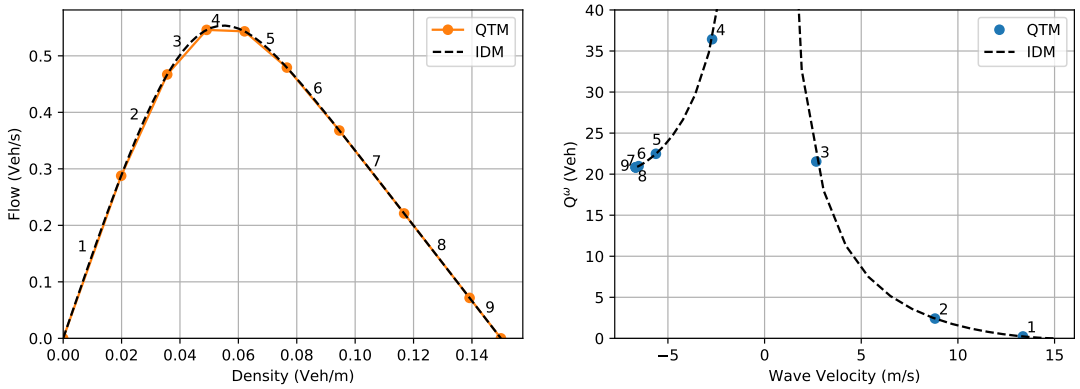


(b)

Figure 5.2: An example of vehicle withholding on a traffic network. (a) shows the flow regions with vehicle withholding in red and the correct flow in green. (b) shows the VT network enumerated with all the possible shortest paths in the solution. Nodes along the first intersection through which a shortest path passes are circled green in both (a) and (b). It is clear that the withholding only occurs on nodes not on any shortest path, which do not contribute to the objective value.



(a)



(b)

Figure 5.3: Example of a QTMW flow with multiple shockwaves. (a) Left plot shows the density of the flow. Forwards and backwards shockwaves are clearly visible. Right plot shows the contours of the Newell surface for the solution, which represent the vehicle trajectories predicted by the LWR model. Overlaid are the trajectories of an IDM microsimulation of the policy which align very well with the QTM prediction. (b) Left plot shows the fundamental diagram used for the QTM solver with 9 piecewise linear segments approximating the IDM fundamental diagram. Right plot shows the values of Q_i^ω, corresponding to each wave u^ω. The wave number ω is enumerated on both plots.

5.5 QTM with Multiple Waves

The results in Sections 5.2 and 5.3 suggest an extension to QTM to model LWR traffic flows with multiple shock-waves. MILP formulations of CTM, LTM and VT such as [Lin and Wang, 2004; Hajiahmadi et al., 2012; Wada et al., 2017] are limited to two wave speeds, and the waves speeds are limited to the slopes of edges within a homogeneous wave lattice. However, we can make use of QTM's non-homogeneous property to model multiple waves with arbitrary speed.

First consider the wave speeds of a fundamental diagram, $Q(k)$. If $Q(k)$ is concave and piece-wise linear, then the slope of each segment h has an associated wave speed u^h and $u^h \geq u^{h+1}$. Let Ω be the set of wave index's for h. We define a set of new variables $q^\omega_{i,n}$ for each $\omega \in \Omega$, and each queue i. Associated with each ω is the constant T^ω_i which defines the propagation time for wave ω at speed u^ω along the length of i. Additionally, there is a constant Q^ω_i associated with each u^ω which gives a bound on the range of $q^\omega_{i,n}$, such that for forward waves, where $u^h > 0$), $q^\omega_{i,n} \in [-\mathrm{Q}^\omega_i, \mathrm{Q}_i]$, and for backward waves, where $u^h < 0$), $q^\omega_{i,n} \in [0, \mathrm{Q}^\omega_i]$, where $\mathrm{Q}^\omega_i = \mathrm{T}^\omega_i \max_k(Q(k) - ku^\omega)$. We may ignore any segments where $u^\omega = 0$, since such segments represent the capacity flow $\max_k Q(k) = \mathrm{F}_{i,j}$ and are modelled by constraint (Q2). Note that for the first segment in Ω, $u^1 = v$, the free flow velocity of the traffic flow, and so $\mathrm{T}^1_i = \mathrm{T}^\mathrm{P}$, and for the final segment in Ω, $\mathrm{Q}^\omega_i = \mathrm{Q}_i$, respectively the queue propagation constant and queue capacity constant of QTM.

To update each $q^\omega_{i,n}$, at each time interval, we replace constraints (Q5) and (Q6) with (Q23) for forward travelling waves, and (Q24) for backward travelling waves.

$$q^\omega_{i,n} = q^\omega_{i,n-1} + V_i(t_{n-1} - \mathrm{T}^\omega_i, t_n - \mathrm{T}^\omega_i) - q^\mathrm{out}_{i,n-1} \qquad \text{for } u^\omega > 0 \qquad \text{(Q23)}$$

$$q^\omega_{i,n} = q^\omega_{i,n-1} + q^\mathrm{in}_{i,n-1} - V^\mathrm{out}_i(t_{n-1} - \mathrm{T}^\omega_i, t_n - \mathrm{T}^\omega_i) \qquad \text{for } u^\omega < 0 \qquad \text{(Q24)}$$

The function V^out in (Q24) is defined in (5.55), and similar to (3.2), represents the volume of traffic exiting i during the time period x to y.

$$V^\mathrm{out}_i(x, y) = (t_{m+1} - x)\frac{q^\mathrm{out}_{i,m}}{\Delta t_m} + \left(\sum_{k=m+1}^{w-1} q^\mathrm{out}_{i,k}\right) + (y - t_w)\frac{q^\mathrm{out}_{i,w}}{\Delta t_w} \tag{5.55}$$

We refer to the MILP (O6, Q1–Q4, Q23, Q24) as the Queue Transmission Model with Waves, (QTMW).

Figure 5.3 shows an example of a solution found using QTMW. An approximation of the IDM fundamental diagram was made using 9 piece-wise linear segments, which in turn were used to generate the values of Q^ω_i for the solver (lower left and right plots). After solving, the Newell surface was generated using the solutions cumulative arrival and departure curves for each link. The density of the solution was found by taking the partial derivative in the negative x direction and this is plotted in the upper left plot. Backwards and forwards travelling shock-waves can be clearly seen as changes in density. Also visible is the spreading out of the traffic flow along the link. The contours of the Newell surface are plotted in the upper right plot, with a contour interval of 1. Each curve represents a vehicle trajectory predicted by the LWR model. Overlaid are the vehicle trajectories of an IDM micro-simulation of the

same signal control policy, after adjustment for lost time. The micro-simulation aligns very well with the QTM solutions prediction.

QTMW as a Constrained Shortest Path Problem

To show that QTMW is a VT shortest path problem, we need to connect the QTMW constraints (Q23) and (Q24) to the VT constraints (5.57) and (5.58), and constraint (Q2) to (5.59).

$$\max \sum_{n=1}^{\mathrm{N}} \sum_{i=1}^{|\mathcal{Q}|} N(x_i^{\text{out}}, t_n) \tag{5.56}$$

$$\text{subject to:} \quad N(x_i^{\text{out}}, t_n) \leq N(x_i^{\text{in}}, t_n - \mathrm{T}_i^{\omega}) + \mathrm{Q}_i^{\omega} \qquad \text{for } u^{\omega} > 0 \tag{5.57}$$

$$N(x_i^{\text{in}}, t_n) \leq N(x_i^{\text{out}}, t_n - \mathrm{T}_i^{\omega}) + \mathrm{Q}_i^{\omega} \qquad \text{for } u^{\omega} < 0 \tag{5.58}$$

$$N(x_i^{\text{out}}, t_n) \leq N(x_i^{\text{out}}, t_{n-1}) + \Delta t_{n-1} \mathrm{F}_{i,j} p_{i,j,n-1} \tag{5.59}$$

Without loss of generality we assume there is an epoch at t_1 where for all i, $q_{i,1}^{\omega} = 0$. To ensure that the VT network is fully connected, Δt_n must be homogeneous and there must be a positive integer $w = {}^{\mathrm{T}_i^{\omega}}/_{\Delta t_n} \in \mathbb{Z}^+$ such that $t_{n-w} = t_n - \mathrm{T}_i^{\omega}$. We can use w to rewrite QTMW constraints (Q23) and (Q24) as (5.60) (5.61), and connect the cumulative counts at nodes $(x_i^{\text{out}}, t_n - \mathrm{T}_i^{\omega})$ and (x_i^{out}, t_n) to the output of queue i with (5.62) and (5.63), and the cumulative counts at nodes $(x_i^{\text{in}}, t_n - \mathrm{T}_i^{\omega})$ and (x_i^{in}, t_n) to the input of queue i with (5.64) and (5.65).

$$q_{i,n}^{\omega} = q_{i,n-1}^{\omega} + q_{i,n-w-1}^{\text{in}} - q_{i,n-1}^{\text{out}} \qquad \text{for } u^{\omega} > 0 \tag{5.60}$$

$$q_{i,n}^{\omega} = q_{i,n-1}^{\omega} + q_{i,n-1}^{\text{in}} - q_{i,n-w-1}^{\text{out}} \qquad \text{for } u^{\omega} < 0 \tag{5.61}$$

$$N(x_i^{\text{out}}, t_n - \mathrm{T}_i^{\omega}) = \sum_{k=1}^{n-w-1} q_{i,k}^{\text{out}} \tag{5.62}$$

$$N(x_i^{\text{out}}, t_n) = \sum_{k=1}^{n-1} q_{i,k}^{\text{in}} \tag{5.63}$$

$$N(x_i^{\text{in}}, t_n - \mathrm{T}_i^{\omega}) = \sum_{k=1}^{n-w-1} q_{i,k}^{\text{in}} \tag{5.64}$$

$$N(x_i^{\text{in}}, t_n) = \sum_{k=1}^{n-1} q_{i,k}^{\text{in}} \tag{5.65}$$

To derive the forwards wave constraint, we rearrange (5.60), and take the cumulative summation over both sides from t_1, to get (5.66). Expanding and simplifying (5.66), we get

(5.68).

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} = \sum_{k=1}^{n-1} \left(q_{i,k}^{\omega} - q_{i,k+1}^{\omega} + q_{i,k-w}^{\text{in}} \right) \tag{5.66}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} = q_{i,1}^{\omega} - q_{i,n}^{\omega} + \sum_{k=1}^{n-w-1} q_{i,k}^{\text{in}} \tag{5.67}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} = -q_{i,n}^{\omega} + \sum_{k=1}^{n-w-1} q_{i,k}^{\text{in}} \qquad (\text{since } q_{i,1}^{\omega} = 0) \tag{5.68}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \leq \mathrm{Q}_i^{\omega} + \sum_{k=1}^{n-w-1} q_{i,k}^{\text{in}} \qquad (\text{since } -q_{i,n}^{\omega} \leq \mathrm{Q}_i^{\omega} \text{ when } u^{\omega} > 0) \tag{5.69}$$

Since $-q_{i,n}^{\omega} \leq \mathrm{Q}_i$, we can relax the equality in (5.68) to get (5.69), which we can see by substituting (5.65) and (5.62) is equivalent to the forwards wave path constraint (5.57). To derive the backwards wave constraint, we rearrange (5.61), and take the cumulative summation over both sides from t_1, to get (5.70). Expanding and simplifying (5.70), we get (5.72).

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{in}} = \sum_{k=1}^{n-1} \left(q_{i,k+1}^{\omega} - q_{i,k}^{\omega} + q_{i,k-w}^{\text{out}} \right) \tag{5.70}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{in}} = q_{i,n}^{\omega} - q_{i,1}^{\omega} + \sum_{k=1}^{n-w-1} q_{i,k}^{\text{out}} \tag{5.71}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{in}} = q_{i,n}^{\omega} + \sum_{k=1}^{n-w-1} q_{i,k}^{\text{out}} \qquad (\text{since } q_{i,1}^{\omega} = 0) \tag{5.72}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{in}} \leq \mathrm{Q}_i^{\omega} + \sum_{k=1}^{n-w-1} q_{i,k}^{\text{out}} \qquad (\text{since } q_{i,n}^{\omega} \leq \mathrm{Q}_i^{\omega} \text{ when } u^{\omega} < 0) \tag{5.73}$$

Since $q_{i,n}^{\omega} \leq \mathrm{Q}_i$, we can relax the equality in (5.72) to get (5.73) which we can see by substituting (5.65) and (5.62) is equivalent to the backwards wave path constraint (5.58). Finally we can use the QTMW constraints (O6, Q1–Q4, Q23, Q24) and the relationships $-q_{i,n}^{\omega} \leq \mathrm{Q}_i^{\omega}$ when $u^{\omega} > 0$, $q_{i,n}^{\omega} \leq \mathrm{Q}_i^{\omega}$ when $u^{\omega} < 0$, and $q_{i,n}^{\text{out}} \leq \Delta t \mathrm{F}_i p_{\ell,n}$, to derive constraints (5.30) to (5.32), where m is the index such that $t_{n-m} = t_n - \mathrm{T}_i^{\mathrm{p}}$.

$$\max \sum_{i=1}^{|\mathcal{Q}|} \sum_{n=1}^{\mathrm{N}} (\mathrm{T} - t_n)\, \Delta t_n f_{i,n}^{\text{out}} \tag{5.74}$$

$$\text{subject to:} \quad \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \leq \mathrm{Q}_i^{\omega} + \sum_{k=1}^{n-w-1} q_{i,k}^{\text{in}} \tag{5.75}$$

$$\sum_{k=1}^{n-1} q_{i,k}^{\text{in}} \leq \mathrm{Q}_i^{\omega} + \sum_{k=1}^{n-w-1} q_{i,k}^{\text{out}} \tag{5.76}$$

$$\sum_{k=1}^{n} q_{i,k}^{\text{out}} \leq \Delta t \mathrm{F}_i p_{\ell,n} + \sum_{k=1}^{n-1} q_{i,k}^{\text{out}} \tag{5.77}$$

The MILP given by constraints (5.74) to (5.77) demonstrates that QTMW with homogeneous

time steps and no turning traffic is solving a constrained shortest path problem.

5.6 Extending VT with turns

In Section 5.3, QTM, QTMW, CTM and LTM were shown to be a equivalent to VT as a constrained shortest path problem on a acyclic directed graph of wave paths, but with the limitation imposed by the VT formulation that the traffic flow be free of turning movements. This suggests a framework to extend VT Theory to traffic networks *with* turning movements, using QTM.

5.6.1 Adding turns with QTM

We can extend VT theory to include turns by splitting the VT graph into complete sub-graphs for each link. Each link can then be solved as a QTM LP and the input and output flows used to connect the sub-graphs. To do so, we add input nodes along the input boundary of each link, along with connecting signal paths. We bind the input nodes to the output nodes of the upstream links, via the sending and receiving flows weighted by the turn probabilities. The input and output flows during each time interval are derived by taking the difference in cumulative counts across the signal paths.

$$\max \sum_{j \in \mathcal{V}_{\text{exit}}} N_j^{\text{out}} \tag{5.78}$$

$$\begin{aligned}
\text{s.t.} \quad N_j^{\text{out}} &\leq N_i^{\text{in}} + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{v}} \cup \mathcal{L}_{\text{d}} \\
N_j^{\text{in}} &\leq N_i^{\text{out}} + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{w}} \\
N_j^{\text{out}} &\leq N_i^{\text{out}} + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{s}} \\
N_j^{\text{in}} &\leq N_i^{\text{in}} + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{s}} \\
q_i^{\text{in}} &= \sum_k \text{Pr}_{ki} q_k^{\text{out}} \quad \forall i \in \mathcal{V} \\
q_j^{\text{out}} &= N_j^{\text{out}} - N_i^{\text{out}} \quad \forall (i,j) \in \mathcal{L}_{\text{s}} \\
q_j^{\text{in}} &= N_j^{\text{in}} - N_i^{\text{in}} \quad \forall (i,j) \in \mathcal{L}_{\text{s}} \\
N_{\text{o}} &= 0
\end{aligned}$$

Where:

i	node at location (t_i, x_i) along a link
N_i^{in}	cumulative vehicle count at link input node i
N_i^{out}	cumulative vehicle count at link output node i
N_{o}	virtual dummy node
$\mathcal{V}$	set of all nodes, including N_{o}
$\mathcal{L}$	set of wave paths, (i, j), the path from i to j
$\mathcal{L}_{\text{v}}$	subset of forward wave paths with slope v
$\mathcal{L}_{\text{w}}$	subset of backward wave paths with slope $-w$
$\mathcal{L}_{\text{s}}$	subset of signal wave paths with slope 0
$\mathcal{L}_{\text{d}}$	subset of dummy node wave paths, (o, i)
q_i^{out}	is the link input flow past i during interval $(i, j) \in \mathcal{L}_{\text{s}}$
q_i^{in}	is the link output flow past i during interval $(i, j) \in \mathcal{L}_{\text{s}}$
$\Pr_{ij}$	turn probability from i to j
c_{ij}	cost from i to j

5.6.2 Stochastic Shortest Path LP

We can simplify the LP (5.78) further by eliminating the input and output flow variables, and directly connecting the sub-graphs for each link. Define N_i^{in} and N_i^{out} as sums over q_i^{in} and q_i^{out}:

$$\begin{aligned} N_j^{\text{in}} &= N_i^{\text{in}} + q_i^{\text{in}} \quad \forall (i,j) \in \mathcal{L}_{\text{s}} \\ N_j^{\text{out}} &= N_i^{\text{out}} + q_i^{\text{out}} \quad \forall (i,j) \in \mathcal{L}_{\text{s}} \\ \implies N_j^{\text{out}} &= \sum_n q_n^{\text{out}} \text{ and } N_j^{\text{in}} = \sum_n q_n^{\text{in}} \quad \forall (n,i) \in \mathcal{L}_{\text{s}} \quad \text{s.t. } t_i < t_j \text{ and } x_i = x_j \end{aligned} \tag{5.79}$$

Then, using the fact that $q_i^{\text{in}} = \sum_k \Pr_{ki} q_k^{\text{out}}$, we can eliminate q_i^{in} and q_i^{out},

$$\begin{aligned} N_i^{\text{in}} &= \sum_n q_n^{\text{in}} \\ &= \sum_n \sum_k \Pr_{ki} q_k^{\text{out}} \\ &= \sum_k \Pr_{ki} \sum_n q_n^{\text{out}} \\ &= \sum_k \Pr_{ki} N_k^{\text{out}} \end{aligned} \tag{5.80}$$

This gives the stochastic shortest path LP:

$$\begin{aligned} & \max \sum_{j \in \mathcal{V}_{\text{exit}}} N_j^{\text{out}} \\ \text{s.t.} \quad & N_j^{\text{out}} \leq N_i^{\text{in}} + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{v}} \cup \mathcal{L}_{\text{d}} \\ & N_j^{\text{in}} \leq N_i^{\text{out}} + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{w}} \\ & N_j^{\text{out}} \leq N_i^{\text{out}} + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{s}} \\ & N_i^{\text{in}} = \sum_k \Pr_{ki} N_k^{\text{out}} \quad \forall i \in \mathcal{V} \\ & N_{\text{o}} = 0 \end{aligned} \tag{5.81}$$

The LP (5.81), shows that solving a QTM traffic flow with turns is equivalent to solving a stochastic shortest path problem.

5.6.3 VT with turns as an Markov Decision Problem value function

Since $N_j^{\text{in}} = \sum_k p_{ki} N_k^{\text{out}}$, we can remove the input nodes and connect the wave path directly to the output nodes at each end. We locate nodes at the exit of each link, just before traffic turns to the downstream links.

$$\begin{aligned} & \max \sum_{j \in \mathcal{V}_{\text{exit}}} N_j \\ \text{s.t.} \quad & N_j \leq N_i + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{s}} \cup \mathcal{L}_{\text{d}} \\ & N_j \leq \sum_k \Pr_{ki} N_i + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{v}} \\ & \sum_k \Pr_{kj} N_j \leq N_i + c_{ij} \quad \forall (i,j) \in \mathcal{L}_{\text{w}} \\ & N_{\text{o}} = 0 \end{aligned} \tag{5.82}$$

Where:

i	node at location (t_i, x_i) along a link
N_i	cumulative vehicle count at node i
N_o	virtual dummy node
$\mathcal{V}$	set of all nodes, including N_o
$\mathcal{L}$	set of wave paths, (i, j), the path from i to j
$\mathcal{L}_\mathrm{v}$	subset of forward wave paths with slope v
$\mathcal{L}_\mathrm{w}$	subset of backward wave paths with slope $-w$
$\mathcal{L}_\mathrm{s}$	subset of signal wave paths with slope 0
$\mathcal{L}_\mathrm{d}$	subset of dummy node wave paths, (o, i)
Pr_{ij}	turn probability from i to j
c_{ij}	cost from i to j

The LP (5.82) shows that finding a traffic flow with turning movements on an extended VT network, is equivalent to finding the optimum value function of a Markov Decision Problem (MDP) [Littman et al., 2013]. By introducing the binary variable p_{ij} to represent the signal state of signal path $(i, j) \in \mathcal{L}_\mathrm{s}$ and substituting $p_{ij}c_{ij}$ as the cost for signal path (i, j), (5.82) becomes a MILP for optimizing signal plans for traffic networks with turn probabilities.

5.7 QTM Continuous Time Solver

The Queue Transmission Model presented in Chapter 3 has the interesting property of piecewise linear dynamics, such that time steps can be non-homogeneous and are only required at change points. This suggests the possibility of an alternate QTM formulation where the time discretization is dynamic and chosen optimally by the solver. Compared to an exogenously defined time discretization, a solution may require significantly less time steps for the same planning horizon, while generating *continuous* time signal schedules of higher fidelity. This section presents such a continuous time QTM solver.

Given a set of queues $\mathcal{Q}$ defining a traffic network, and a planning horizon T, we choose a number of time intervals N_i for each queue $i \in \mathcal{Q}$ and define a vector $\vec{\delta t}_i$, of N_i time step variables, with each $\delta t_{i,n} \in [0, T]$ and $n \in \mathrm{N}_i$. The vector $\vec{\delta t}_i$ then defines a dynamic schedule of time intervals for each queue starting from time 0, with the input schedule given by constraints (W1) and (W2), and the output schedule offset from the input by the travel time T_i^P with constraint (W3). The solver is free to place each $(t_{i,n}^\mathrm{in}, t_{i,n}^\mathrm{out})$ pair exactly as needed to define changes points in the queue dynamics.

$$t_{i,1}^\mathrm{in} = 0 \tag{W1}$$

$$t_{i,n+1}^\mathrm{in} = t_{i,n}^\mathrm{in} + \delta t_{i,n} \tag{W2}$$

$$t_{i,n}^\mathrm{out} = t_{i,n}^\mathrm{in} + \mathrm{T}_i^\mathrm{P} \tag{W3}$$

Since each queue has a unique time discretization, in order to move vehicles between queues we need to find a mapping from upstream to downstream time intervals. If j flows into i, we define a 0–1 incidence matrix $\mathbf{U_i}$ that maps time intervals in j to i. $\mathbf{U_i}$ is an $\mathrm{N}_i \times \mathrm{M}_j$ matrix, where M_j is the number of time intervals in j. A one in entry (n, m) of $\mathbf{U_i}$ indicates that interval $\delta t_{i,n}$ is within interval $\delta t_{j,m}$. In order to find the overlap between $\delta t_{i,n}$ and $\delta t_{j,m}$, we use two binary variables $v_{i,n,m}$ and $z_{i,n,m}$, and a method similar to [Lo, 1998] and [Beard and Ziliaskopoulos, 2006] with constraints (W4) to (W7) and $\mathcal{M}$ and ϵ as sufficiently large and small constants respectively. If $t^{\text{in}}_{j,n} \geq t^{\text{out}}_{j,n}$ then $v_{i,n,m} = 1$ else 0, and if $t^{\text{in}}_{j,n} \leq t^{\text{out}}_{j,n}$ then $z_{i,n,m} = 1$ else 0. Then if $v_{i,n,m} = 1$ and $z_{i,n+1,m+1} = 1$ then the interval $\delta t_{i,n}$ must lie within $\delta t_{j,m}$ and this is enforced with constraint (W8) where $u_{i,n,l}$ is the (n, m) entry in $\mathbf{U}_i$.

$$\mathcal{M}v_{i,n,m} - \epsilon \geq t^{\text{in}}_{i,n} - t^{\text{out}}_{j,m} \tag{W4}$$

$$-\mathcal{M}(1 - v_{i,n,m}) \leq t^{\text{in}}_{i,n} - t^{\text{out}}_{j,m} \tag{W5}$$

$$\mathcal{M}z_{i,n,m} - \epsilon \geq t^{\text{out}}_{j,m} - t^{\text{in}}_{i,n} \tag{W6}$$

$$-\mathcal{M}(1 - z_{i,n,m}) \geq t^{\text{out}}_{j,m} - t^{\text{in}}_{i,n} \tag{W7}$$

$$u_{i,n,m} = v_{i,n,m} + z_{i,n+1,m+1} - 1 \tag{W8}$$

However, we do not require that all intervals in i map to intervals in j – only those intervals carrying traffic flow are needed, and further, depending on the rate of flow, multiple intervals in i can be mapped to an interval in j. Our only constraint is that when there is a mapping, the total duration of the intervals in i should be exactly equal to the duration of the interval they map to in j. We can enforce these requirements by placing constraints on the entries of $\mathbf{U}_i$. Firstly, to detect when there is non zero flow during $\delta t_{j,m}$, we use constraints (W9) and (W10) to set a binary variable $r^{\text{in}}_{i,m} = 1$ for each interval in j, only if $q^{\text{out}}_{j,m}$ is strictly greater than 0 during interval m, otherwise 0. If $r^{\text{in}}_{i,m} = 1$ then we must map interval m in j to at least one interval in i which is represented with at least one element in column m of $\mathbf{U}_i$ set to 1. The condition can be met by summing column m of $\mathbf{U}_i$ to be 1 with constraint (W11).

$$\mathcal{M}(1 - r^{\text{in}}_{i,m}) \geq -q^{\text{out}}_{j,m} + \epsilon \tag{W9}$$

$$\mathcal{M}r^{\text{in}}_{i,m} \geq q^{\text{out}}_{j,m} - \epsilon \tag{W10}$$

$$\sum_n u_{i,n,m} \geq r^{\text{in}}_{i,m} \tag{W11}$$

Further, if the outflow from j during interval m is at the maximum rate F_j, then we are free to map several intervals in i to m, since the number of vehicles entering during each mapped interval in i can be determined simply by $\mathrm{F}_j \delta t_{i,n}$. This condition can be guaranteed only when the value $q_{j,m}$ is strictly greater than zero, since by definition the QTM discharges queues at the maximum rate when there is no downstream restriction. To detect when $q_{i,n}$ is strictly greater than zero, we use constraints (W12) and (W13) to set $w^{\text{out}}_{i,n} = 1$ when $q_{i,n} = 0$, and to zero otherwise. Then we can set $r^{\text{out}}_{i,n} = 0$ when $q_{i,n} > 0$ and the traffic light is green during n as indicated by $p_{i,n}$, using constraints (W14) to (W16). Finally we can set the upper bound on the number of intervals in i mapped to interval m in j, with constraint (W17), which reduces

to at most one element in column m of $\mathbf{U}_\mathrm{i}$ when the outflow from j is less than F_j.

$$\mathcal{M}(1 - w^{\text{out}}_{i,n}) \geq -q_{i,n} + \epsilon \tag{W12}$$
$$\mathcal{M}w^{\text{out}}_{i,n} \geq q_{i,n} - \epsilon \tag{W13}$$
$$1 - p_{i,n} + w^{\text{out}}_{i,n} \geq r^{\text{out}}_{i,n} \tag{W14}$$
$$1 - p_{i,n} \leq r^{\text{out}}_{i,n} \tag{W15}$$
$$w^{\text{out}}_{i,n} \leq r^{\text{out}}_{i,n} \tag{W16}$$
$$\sum_n u_{i,n,m} \leq 1 + \mathcal{M}r^{\text{out}}_{j,m} \tag{W17}$$

If the flow rate is less than F_j then the number of vehicles arriving during interval n depends on the ratio of $\delta t_{i,n}$ to $\delta t_{j,m}$ which cannot be determined using linear constraints. In which case we must map $\delta t_{i,n} = \delta t_{j,m}$. Likewise, if the flow rate is F_j then the total duration of intervals in i mapped to interval m by $\mathbf{U}_i$ must sum to $\delta t_{j,m}$, as given by equation (5.83).

$$\delta t_{j,m} = \sum_n \delta t_{i,n} u_{i,n,m} \tag{5.83}$$

To encode the relationship in (5.83) as constraints, first we construct a helper matrix $\mathbf{A}_i$ using constraints (W18) to (W20) where $a_{i,n,m}$ is element (n, m) of $\mathbf{A}_i$, such that $a_{i,n,m} = \delta t_{i,n} u_{i,n,m}$. Then the of summation of column m in $\mathbf{A}_i$ can be equated with $\delta t_{j,m}$ using constraints (W21) and (W22).

$$a_{i,n,m} \leq \delta t_{i,n} + \mathcal{M}(1 - u_{i,n,m}) \tag{W18}$$
$$a_{i,n,m} \geq \delta t_{i,n} - \mathcal{M}(1 - u_{i,n,m}) \tag{W19}$$
$$a_{i,n,m} \leq \mathcal{M}u_{i,n,m} \tag{W20}$$
$$\delta t_{j,m} \leq \sum_n a_{i,n,m} + \mathcal{M}(1 - r^{\text{in}}_{i,m}) \tag{W21}$$
$$\delta t_{j,m} \geq \sum_n a_{i,n,m} - \mathcal{M}(1 - r^{\text{in}}_{i,m}) \tag{W22}$$

Next we need to find the volume of traffic $q^{\text{in}}_{i,n}$ flowing from queue j into queue i during each interval n. We use a set of intermediate variables $f^{\text{in}}_{i,n,m}$ to hold the values as determined by $u_{i,n,m}$ and $r^{\text{out}}_{j,m}$. If $u_{i,n,m} = 0$ then $f^{\text{in}}_{i,n,m} = 0$ using constraint (W23). If $u_{i,n,m} = 1$ and $r^{\text{out}}_{j,m} = 1$ then the output of j during interval m is not flowing at the maximum rate F_j, $\delta t_{i,n} = \delta t_{j,m}$, and all the outflow from $q^{\text{out}}_{j,m}$ is transferred to $f^{\text{in}}_{i,n,m}$ by constraints (W24) and (W25). Otherwise, if $u_{i,n,m} = 1$ and $r^{\text{out}}_{j,m} = 0$ then $\delta t_{i,n} \leq \delta t_{j,m}$, the output of j during interval m is at the constant rate F_j, and $\mathrm{F}_j \delta t_{i,n} \leq q^{\text{out}}_{j,m}$ vehicles are transferred during interval n using constraints (W26) and (W27). Finally the value of $q^{\text{in}}_{i,n}$ is determined by totaling all of the $f^{\text{in}}_{i,n,m}$ values with

constraint (W28).

$$f^{\text{in}}_{i,n,m} \leq \mathcal{M} u_{i,n,m} \tag{W23}$$
$$f^{\text{in}}_{i,n,m} \leq q^{\text{out}}_{j,m} + \mathcal{M}(1 - r^{\text{out}}_{j,m}) + \mathcal{M}(1 - u_{i,n,m}) \tag{W24}$$
$$f^{\text{in}}_{i,n,m} \geq q^{\text{out}}_{j,m} - \mathcal{M}(1 - r^{\text{out}}_{j,m}) - \mathcal{M}(1 - u_{i,n,m}) \tag{W25}$$
$$f^{\text{in}}_{i,n,m} \leq \mathrm{F}_j \delta t_{i,n} \tag{W26}$$
$$f^{\text{in}}_{i,n,m} \geq \mathrm{F}_j \delta t_{i,n} - \mathcal{M} r^{\text{out}}_{j,m} - \mathcal{M}(1 - u_{i,n,m}) \tag{W27}$$
$$q^{\text{in}}_{i,n} = \sum_{l} f^{\text{in}}_{i,n,l} \tag{W28}$$

Similarly, we need to find the volume of traffic $q^{\text{out}}_{i,n}$ flowing out from queue i during each interval n. Firstly, if the traffic signal controlling i is red during interval n, then $p_{i,n} = 0$ and $q^{\text{out}}_{i,n} = 0$ is enforced with constraint (W29). Otherwise, if the signal is green and $q_{i,n} = 0$ then i is discharging at the same rate is the inflow, and $p_{i,n} = 1$ and $r^{\text{out}} = 1$ sets $q^{\text{out}}_{i,n} = q^{\text{in}}_{i,n}$ with constraints (W30) and (W31). Finally if the signal is green and $q_{i,n} > 0$, then i is discharging at the rate F_i and $p_{i,n} = 1$ and $r^{\text{out}} = 0$ sets $q^{\text{out}}_{i,n} = \mathrm{F}_i \delta t_{i,n}$ using constraints (W32) and (W33).

$$q^{\text{out}}_{i,n} \leq \mathcal{M} p_{i,n} \tag{W29}$$
$$q^{\text{out}}_{i,n} \leq q^{\text{in}}_{i,n} + \mathcal{M}(1 - p_{i,n}) + \mathcal{M}(1 - r^{\text{out}}_{i,n}) \tag{W30}$$
$$q^{\text{out}}_{i,n} \geq q^{\text{in}}_{i,n} - \mathcal{M}(1 - p_{i,n}) - \mathcal{M}(1 - r^{\text{out}}_{i,n}) \tag{W31}$$
$$q^{\text{out}}_{i,n} \leq \mathrm{F}_i \delta t_{i,n} \tag{W32}$$
$$q^{\text{out}}_{i,n} \geq \mathrm{F}_i \delta t_{i,n} - \mathcal{M}(1 - p_{i,n}) - \mathcal{M} r^{\text{out}}_{i,n} \tag{W33}$$

Finally, we can update the state of queue i at the end of interval i using constraint (W34).

$$q_{i,n+1} = q_{i,n} + q^{\text{in}}_{i,n} - q^{\text{out}}_{i,n} \tag{W34}$$

A continuous time traffic light schedule L representing M cycles, can be defined with $\mathrm{M} + 1$ time steps, $t^{\mathrm{L}}_{\ell,m}$, and an $\mathrm{M} \times \mathrm{K}$ incidence matrix, S^{L}. Each column k of S^{L} is a vector of zeros and ones that map intervals in L to phase k. Let $s^{\mathrm{L}}_{m,k}$ be element (m, k) of S^{L}, then $s^{\mathrm{L}}_{m,k} = 1$ if phase k is green during interval m, zero otherwise. To enforce the minimum and maximum duration's of each phase activation, we use constraints (W35) and (W36). Additionaly, if we wish to have a fixed time schedule with optimized phase duration and offset, we can add a variable ϕ^{fixed}_k for each fixed phase k and the optional constraint (W37).

$$t^{\mathrm{L}}_{k,m+1} - t^{\mathrm{L}}_{k,m} \leq \Phi^{\max}_{k,m} \tag{W35}$$
$$t^{\mathrm{L}}_{k,m+1} - t^{\mathrm{L}}_{k,m} \geq \Phi^{\min}_{k,m} \tag{W36}$$
$$t^{\mathrm{L}}_{k,m+1} - t^{\mathrm{L}}_{k,m} = \phi^{\text{fixed}}_k \tag{W37}$$

To determine the value of $p_{i,n}$, the traffic signal controlling queue i during interval n, we can use a method similar to constraints (W4) to (W7) to detect overlapping intervals between $\vec{\delta t}_i$ and each phase, k of L. We use introduce two binary variables $v^{\mathrm{L}}_{i,n,m}$ and $z^{\mathrm{L}}_{i,n,m}$ and use constraints (W38) to (W41) to indicate the overlapping time steps. Then an incidence ma-

trix $\mathbf{U}^{\mathrm{L}}$ can be defined to indicate those intervals of i that fall within an interval of L, with constraint (W42).

$$\mathcal{M}v^{\mathrm{L}}_{i,n,m} - \epsilon \geq t^{\mathrm{out}}_{i,n} - t^{\mathrm{L}}_{k,m} \qquad \text{(for each phase } k \text{ in L)} \tag{W38}$$

$$-\mathcal{M}(1 - v^{\mathrm{L}}_{i,n,m}) \leq t^{\mathrm{out}}_{i,n} - t^{\mathrm{L}}_{k,m} \qquad \text{(for each phase } k \text{ in L)} \tag{W39}$$

$$\mathcal{M}z^{\mathrm{L}}_{i,n,m} - \epsilon \geq t^{\mathrm{L}}_{k,m} - t^{\mathrm{out}}_{i,n} \qquad \text{(for each phase } k \text{ in L)} \tag{W40}$$

$$-\mathcal{M}(1 - z^{\mathrm{L}}_{i,n,m}) \geq t^{\mathrm{L}}_{k,m} - t^{\mathrm{out}}_{i,n} \qquad \text{(for each phase } k \text{ in L)} \tag{W41}$$

$$u^{\mathrm{L}}_{i,n,l} = v^{\mathrm{L}}_{i,n,l} + z^{\mathrm{L}}_{i,n+1,l+1} - 1 \tag{W42}$$

However we don't need to map every interval of $\vec{\delta}t_i$ to within an interval of L. Only those intervals with either traffic flowing or where $q_{>,i}0$ and the phase of L controlling i is red need to be detected. We already have constraints (W12) and (W13) to set $w^{\mathrm{out}}_{i,n} = 0$ when $q_{i,n} > 1$, and we define a binary variable $w^{\mathrm{in}}_{i,n} = 1$ when $q^{\mathrm{in}}_{i,n} > 0$, zero otherwise, using constraints (W43) and (W44).

$$\mathcal{M}(1 - w^{\mathrm{in}}_{i,n}) \geq -q^{\mathrm{in}}_{i,n} + \epsilon \tag{W43}$$

$$\mathcal{M}w^{\mathrm{in}}_{i,n} \geq q^{\mathrm{in}}_{i,n} - \epsilon \tag{W44}$$

Now we can set element $u^{\mathrm{L}}_{i,n,m}$ of $\mathbf{U}^{\mathrm{L}}$ to 1 if either $w^{\mathrm{in}} = 1$ or $w^{\mathrm{out}} = 1$, or to zero otherwise, with constraints (W45) and (W46), and $p_{i,n}$ is set to 0 or 1 by the sum of column m of $\mathbf{U}^{\mathrm{L}}$ weighted by column k of $\mathbf{S}^{\mathrm{L}}$ using constraint (W47).

$$\sum_l u^{\mathrm{L}}_{i,n,m} \geq 1 - w^{\mathrm{in}}_{i,n} \tag{W45}$$

$$\sum_l u^{\mathrm{L}}_{i,n,m} \geq 1 - w^{\mathrm{out}}_{i,n} \tag{W46}$$

$$p_{i,n} = \sum_m s^{\mathrm{L}}_{k,m} u^{\mathrm{L}}_{i,n,m} \tag{W47}$$

Finally, we want to force the solver to use all the N_i intervals available to cover as much of the planning horizon T as possible. We define an objective term to minimizing the absolute value of the difference between T and the end of the last time interval of each queue, $t^{\mathrm{out}}_{i,\mathrm{N}_i+1}$. We add a new variable $d^{\mathrm{T}}_i \geq 0$ for each queue i and set $d^{\mathrm{T}}_i = \left|t^{\mathrm{out}}_{i,\mathrm{N}_i+1} - \mathrm{T}\right|$ using constraints (W48) and (W49).

$$d^{\mathrm{T}}_i \geq t^{\mathrm{out}}_{i,\mathrm{N}_i+1} - \mathrm{T} \tag{W48}$$

$$d^{\mathrm{T}}_i \geq -t^{\mathrm{out}}_{i,\mathrm{N}_i+1} + \mathrm{T} \tag{W49}$$

Now we can set the objective to minimize each d^{T}_i with constraint (O7).

$$\min \sum_i d^{\mathrm{T}}_i \tag{O7}$$

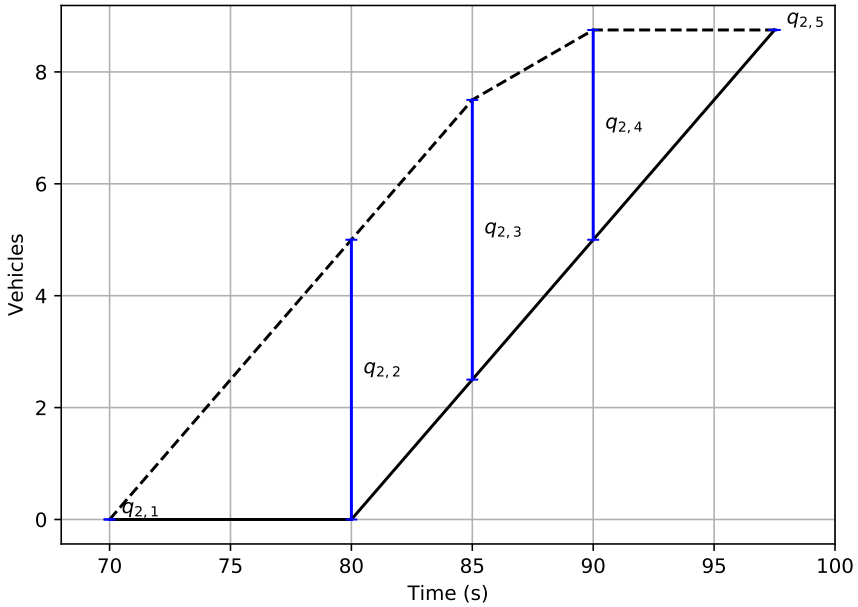


Figure 5.4: Example of a cumulative departure curve for q_2. iThe q_2 values are exactly the difference between departures without delay (dashed line) and the actual departures with delay. The area between the curves is the delay in vehicle seconds.

5.7.1 Objective Functions

Since the time intervals of the solver are also variables, we cannot formulate a objective function to minimize total travel time or delay in a way similar to (O6). However we know that the value of the queue variable gives the vertical distance between the departures without delay and the actual departures with delay, and together with Δt gives the area which is the delay in vehicle seconds (see Fig. 5.4). The area of the trapezoid formed between the two curves is given by (5.84) for each interval n and the sum of all such area gives the total delay in the network. Since $\delta t_{i,n}$ is a variable we cannot calculate the area directly with (5.84). However, during each interval the inflow and outflow is constant in QTM, therefore we can derive (5.87) for the area without directly using time, and if $f_{i,n}^{\text{in}} \neq f_{i,n}^{\text{out}}$ then the area is proportional to $q_{i,n}$.

$$A_n = \delta t_{i,n} \frac{q_{i,n} + q_{i,n+1}}{2} \tag{5.84}$$

$$q_{i,n+1} + \delta t_{i,n} f_{i,n}^{\text{in}} = q_{i,n} + \delta t_{i,n} f_{i,n}^{\text{out}} \tag{5.85}$$

$$\delta t_{i,n} = \frac{q_{i,n} - q_{i,n+1}}{f_{i,n}^{\text{out}} - f_{i,n}^{\text{in}}} \tag{5.86}$$

$$A_n = \frac{(q_{i,n} - q_{i,n+1})(q_{i,n} + q_{i,n+1})}{2(f_{i,n}^{\text{out}} - f_{i,n}^{\text{in}})} \tag{5.87}$$

$$A_n \propto q_{i,n} \quad \text{if } f_{i,n}^{\text{out}} \neq f_{i,n}^{\text{in}} \tag{5.88}$$

This would suggest that minimizing each $q_{i,n}$ in the objective would have the effect of minimizing delay in the network, except during those intervals where $f^{\text{out}}_{i,n} = f^{\text{in}}_{i,n}$ (for example, the region between $q_{2,2}$ and $q_{2,3}$ in Fig. 5.4). We could detect those intervals where $f^{\text{out}}_{i,n} = f^{\text{in}}_{i,n}$ and derive a term for the objective function. However, simply minimizing the duration of all intervals where $q_{i,n} > 0$, is more straightforward, and we already have the variable $w^{\text{out}}_{i,n}$ to indicate the condition. To minimize the duration of intervals where $q_{i,n} > 0$, first we define a new variable $d^{\text{q}}_{i,n}$ and use $w^{\text{out}}_{i,n}$ in constraints (W50) and (W51) to set $d^{\text{q}}_{i,n} = \delta t_{i,n}$ only when $q_{i,n} > 0$.

$$d^{\text{q}}_{i,n} \leq \delta t_{i,n} + \mathcal{M} w^{\text{out}}_{i,n} \tag{W50}$$

$$d^{\text{q}}_{i,n} \geq \delta t_{i,n} - \mathcal{M} w^{\text{out}}_{i,n} \tag{W51}$$

Finally, adding terms for $q_{i,n}$ and $d^{\text{q}}_{i,n}$ to the objective (O7) gives objective (O8) to minimize the total delay in the network.

$$\min \sum_i d^{\text{T}}_i + \sum_i \sum_n q_{i,n} + \sum_i \sum_n d^{\text{q}}_{i,n} \tag{O8}$$

[Lin and Wang, 2004] show that the difference between $y_{i,n+1}$ and $y_{i,n}$ is a good approximation of the number of stops. We can derive a similar objective function using $|q_{i,n+1} - q_{i,n}|$. We introducing a new variable $d^{\text{stop}}_{i,n} > 0$ and set $d^{\text{stop}}_{i,n} = |q_{i,n+1} - q_{i,n}|$ using constraints (W52) and (W53) to give an alternative objective function (O9) to minimize the number of stops.

$$q_{i,n+1} - q_{i,n} \leq d^{\text{stop}}_{i,n} \tag{W52}$$

$$-q_{i,n+1} + q_{i,n} \leq d^{\text{stop}}_{i,n} \tag{W53}$$

$$\min \sum_i d^{\text{T}}_i + \sum_i \sum_n d^{\text{stop}}_{i,n} \tag{O9}$$

5.7.2 An illustrative example

An example of a simple network with two queues controlled by two traffic signals, solved using the MILP (W1) – (W49), (O7) is shown Fig. 5.5. In the example $N_1 = 4$ and $N_2 = 4$, and

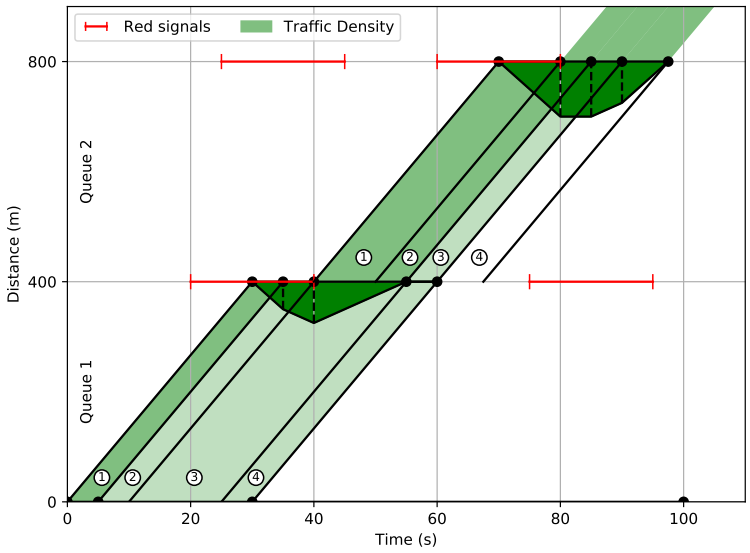


Figure 5.5: QTM Continuous solution example. Only 4 samples per queue are required to find the *exact* solution. Intervals 1 and 2 in queue 2 are mapped to interval 3 in queue 1, and interval 3 in queue 2 is mapped to interval 4 in queue 1. Interval 4 in queue 2, needed to discharge queue 2, is not mapped to any interval in queue 1.

the state of the variables in the solution is shown below.

$$\vec{\delta t}_1 = \begin{bmatrix} 5 & 5 & 15 & 5 \end{bmatrix} \tag{5.89}$$

$$\vec{\delta t}_2 = \begin{bmatrix} 10 & 5 & 5 & 7.5 \end{bmatrix} \tag{5.90}$$

$$\mathbf{U}_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.91}$$

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 0 & 10 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 5 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.92}$$

$$\vec{w}_1^{\text{out}} = \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix} \tag{5.93}$$

$$\vec{r}_1^{\text{out}} = \begin{bmatrix} 1 & 1 & 0 & 1 \end{bmatrix} \tag{5.94}$$

$$\vec{r}_2^{\text{in}} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix} \tag{5.95}$$

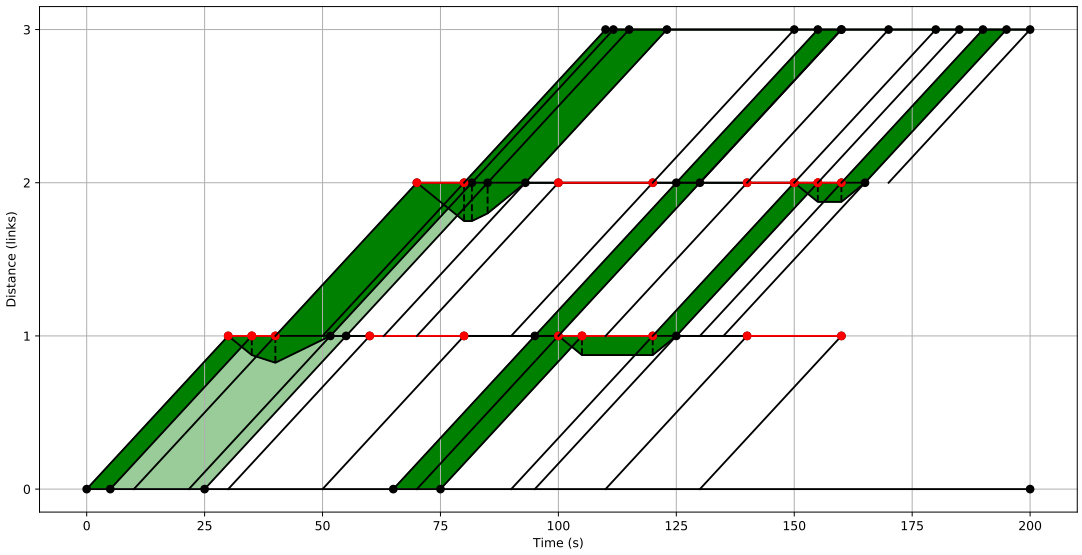
From the entries in the $\mathbf{U}_1$ we can see that the third interval in q_1, $\delta t_{1,3}$, is mapped to $\delta t_{2,1}$ and $\delta t_{2,2}$, in q_2, and $\delta t_{1,4}$ in q_1 is mapped to $\delta t_{2,3}$ in q_2. Furthermore, the entries in the $\mathbf{A}_1$ force the sum of the durations of $\delta t_{2,3}$ and $\delta t_{2,4}$ to equal the duration of $\delta t_{1,3}$, and that $\delta t_{1,4} = \delta t_{2,3}$. The entries in $\vec{w}_1^{\text{out}}$ indicate that q_2 is zero at the start of intervals 1 and 4, and $\vec{r}_1^{\text{out}}$ indicates that q_2 is discharging at a rate less than than F_1 during intervals 1,2 and 4. Finally, $\vec{r}_2^{\text{in}}$ indicates

that q_2 is receiving inflow at the rate F_1 during intervals 1 and 2. The cumulative departure curve for q_2 is shown in Fig. 5.4.

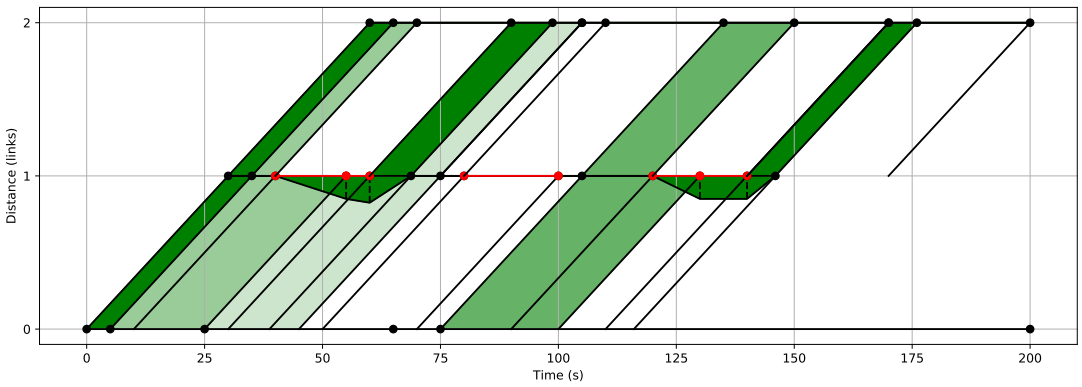
An extended example of a QTM continuous time solution are shown in Fig. 5.6 for an arterial with two intersections (Fig. 5.6(a)), and a side street (Fig. 5.6(b)). The solver is allocated 13 intervals per queue to find a solution, however, not all intervals are needed in all the queues. Some of the intervals are given zero duration, while others are mapped arbitrarily, and contribute nothing to the solution.

5.8 Summary

In this chapter, we showed that with homogeneous time steps and no turning traffic, MILP formulations for QTM, CTM, and LTM, are equivalent to a VT MILP for finding LWR traffic flows as a constrained shortest path problem on a VT wave lattice. QTM was further extended to model multiple waves, and showed that traffic withholding does not impact on the solution if the withholding term is removed from the objective function (or equivalently, $\beta = 0$). We then used QTM to extend VT theory to include turning movements, and showed that the problem of optimizing signal plans on traffic networks with turning movements, is equivalent to finding the optimal MDP value function. Finally, we introduced a continuous time MILP to solve QTM traffic flows exactly, and described an objective function for minimizing delay in signal optimization using the MILP.



(a)



(b)

Figure 5.6: An extended QTM continuous time solution example over several intersections along an arterial (a) with a side street (b). The solver is given 13 intervals per queue. Not all the intervals are needed by all the queues to find the solution. Some of the unused intervals are given zero duration, while others are mapped arbitrarily by the solver.

Chapter 6

QTM Comparisons with CTM, LTM and VT

6.1 Overview

In Chapter 5 we showed that QTM, CTM and LTM MILPS are all equivalently solving a VT shortest path problem. The question arises how well do these models perform both in terms of solution quality and solve time? In this chapter we compare QTM with the CTM, LTM and VT traffic MILP's. We run a series of experiments to compare the relative performance of all these models, using a range of traffic networks and demand scenarios.

6.2 Empirical Evaluation

In this section the performance of several MILP based traffic formulations is compared across a range of traffic networks and demand scenarios. We consider the scalability by comparing the CPU time required to solve the MILP, and solution quality in terms of delay distribution.

6.2.1 Models

We compare the following MILP formulations:

- QTM as defined by (O6, Q1–Q6).
- CTM as defined by (O2, C1–C4, C8, C10, C12, C24, C26).
- LTM as defined by (O4, L1–L14).
- VT as defined by (O4, V2).

In combination with the four traffic flow models above, we evaluate three signal timing models:

- The moving window method of [Lin and Wang, 2004], referred to here as Lin, and defined as (C19–C22).
- QTM, defined as (Q7–Q16).
- The Signal Constraint Network of [Wada et al., 2017], referred to here as SCN and defined as (V3, V4).

The features of the models are listed in Table 6.1. Due to the heterogeneous nature of the features a full comparison is not possible. For example, The formulations of [Wada et al., 2017] and [Lin and Wang, 2004] do not support turning movements, so we must limit our evaluation to road networks without turns, and the formulation of [Hajiahmadi et al., 2012] provides no signal constraints, so we must extend the formulation with one of the other models signal constraints. The formulations of [Wada et al., 2017] has no constraints for maximum green, so we can not bound green duration in the evaluations.

Feature Matrix					
MILP Formulation	Traffic Model	Time Discretization	Turn Movements	Lost Time	Signal Constraints
Guilliard et al. [2020]	QTM	Non-homogeneous	Yes	Yes	Yes
Wada et al. [2017]	VT	Homogeneous	No	Yes	Yes[1]
Lin and Wang [2004]	CTM	Homogeneous	No	No	Yes
Hajiahmadi et al. [2012]	LTM	Homogeneous	Yes	No	None

[1] No maximum green constraints.

Table 6.1: The features of the models under comparison. The heterogeneous nature of the features makes a full comparison challenging.

6.2.2 Network Parameters

To measure the scalability of the models we use rectangular grid networks ranging in size from one intersection (1×1), to 81 intersections (9×9). To evaluate the impact of intersection spacing on the signal optimization problem, we generate two versions of each grid: one with fixed spacing of 417 m between intersections, and another where the spacing between intersections varies. To maintain the rectangular structure of the grid, we do this by varying the spacing between each row and each column of the grid between 139 m and 417 m. However, since the CTM, LTM and VT formulations require the link duration to be an integer multiple of the time step, the spacings are chosen from one of three values, 139 m, 278 m, or 417 m. This corresponds to the three possible link travel time durations of 10, 20 or 30 s when the free flow speed is 13.89 m/s and the time step is 10 s. To further increase the difficulty, for each grid we use two types of street layout, one-way streets and two-way streets. See Fig. 6.1 for an example of a 9×9 grid with two-way streets and varying spacing.

The fundamental diagram is assumed to be the same for each link, and uses the triangular fundamental diagram in Fig. 6.2(a).

Each intersection is controlled by a traffic light with two phases: east-west and north-south, $\mathcal{P}_\ell = \{EW_\ell, NS_\ell\}$. Since some of the formulations under evaluation have no maximum green time or cycle time constraints, no such constraints are placed on the signal timing. Whenever lost time is considered, a duration of 10 s is used ($T_\ell^{\mathrm{L}} = 10$s for QTM) made up of 6 s of start up lost time, 2 seconds of yellow lost time and 2 s of all red.

6.2.3 Demand Profiles

Each network is evaluated by applying demand profiles to the inputs, consisting of platoons with varying length and arrival time. Platoons enter the network with a flow rate of 0.5 vehicles/s and density of 0.036 vehicles/m (the peak of the triangular fundamental diagram in Fig. 6.2(a)), and each vehicle is travelling at 13.89 m/s. Platoons are injected into the network until $t = 200$ s, after which the inflow rate is set to zero. Since the arrival times and durations of the platoons varies between the inputs, a more challenging problem is created for the solver to find an optimal signal plan. Predicting such demand profiles seems reasonable, since existing vehicle sensor technology such as induction loops, radar, and cameras have been used to detect platoons approaching intersections, and estimate their length and arrival

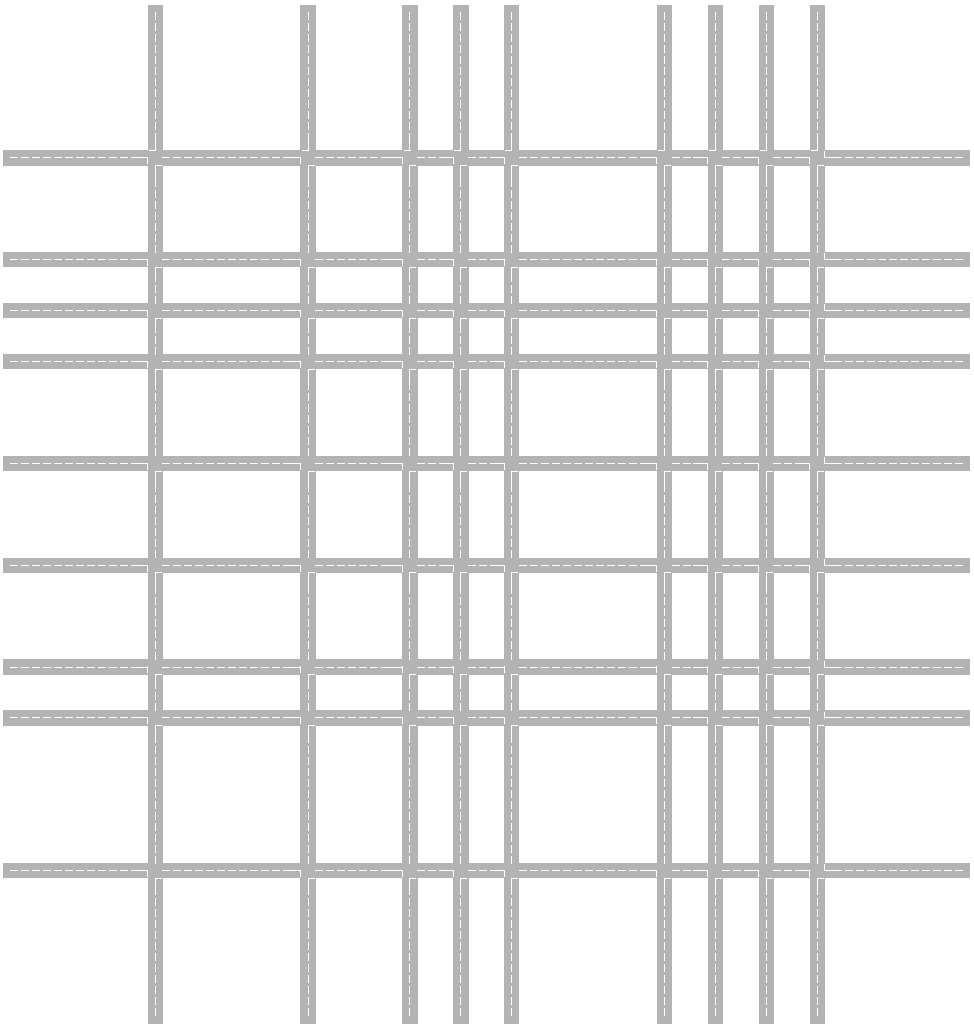
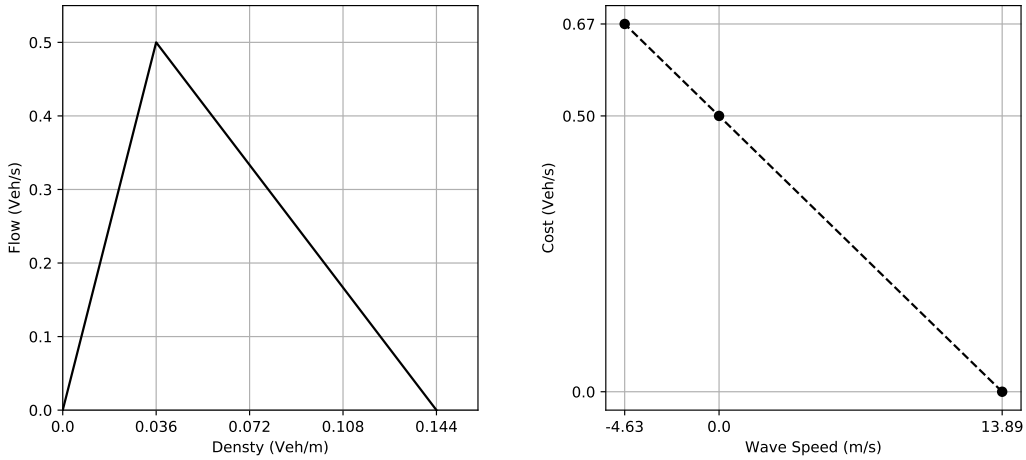


Figure 6.1: An example of the largest network used in the experiments with 81 signal controlled intersections, and two-way streets. The distance between intersections varies between 139 m and 417 m.



(a) Fundamental Diagram

(b) Wave cost function

Figure 6.2: (a) The triangular fundamental diagram used for all links in the comparison experiments. The forward wave-speed is 13.89 m/s, the backwards wave-speed is 4.63 m/s, the maximum flow rate of 0.5 Vehicles/s at a density of 0.036 Vehicles/m and the jam density is 0.144 Vehicles/m. (b) The wave cost function corresponding to the fundamental diagram in (a), gives the path costs used in the VT formulations' lopsided network.

time [Smith et al., 2013].

6.2.4 Model Parameters

For all models, we use a time step $\Delta t = 10$ s, and a free flow speed of $v_0 = 13.89$ m/s. From the triangular fundamental diagram in Fig. 6.2(a) we get a forward wave speed of $v = 13.89$ m/s, a backwards wave speed of $w = 4.63$ m/s. The maximum flow rate $\mathrm{F} = 0.5$ and a jam density $\rho^{\mathrm{jam}} = 0.144$ veh/m.

QTM

For each queue i, the values for propagation time, $\mathrm{T}_i^{\mathrm{p}} = \frac{L}{v_0} \in \{10\,\mathrm{s}, 20\,\mathrm{s}, 30\,\mathrm{s}\}$ and the queue capacity $\mathrm{Q}_i = \rho^{\mathrm{jam}} L \in \{20\,\mathrm{veh}, 40\,\mathrm{veh}, 60\,\mathrm{veh}\}$, where $L \in \{139\,\mathrm{m}, 278\,\mathrm{m}, 417\,\mathrm{m}\}$ is the link length. The maximum flow rate between upstream queue i and downstream queue j is set at $\mathrm{F}_{i,j} = 0.5$ veh/s, and since there are no turning movements $\mathrm{Pr}_{i,j} = 1$. For output links $\mathrm{F}_i^{\mathrm{out}} = 0.5$ veh/s and zero for all other links, and for input links the inflow matrix $\mathbf{I}$ is determined from the generated demand profiles. Each non-output queue is controlled by a traffic single phase k of light ℓ, where $k \in \{EW_\ell, NS_\ell\}$, depending on the link orientation east-west or north-south. In order to evaluate the impact of withholding on the solution, we use two different values of β for the objective function weight in (O6), $\beta = 0.001$ and $\beta = 0$.

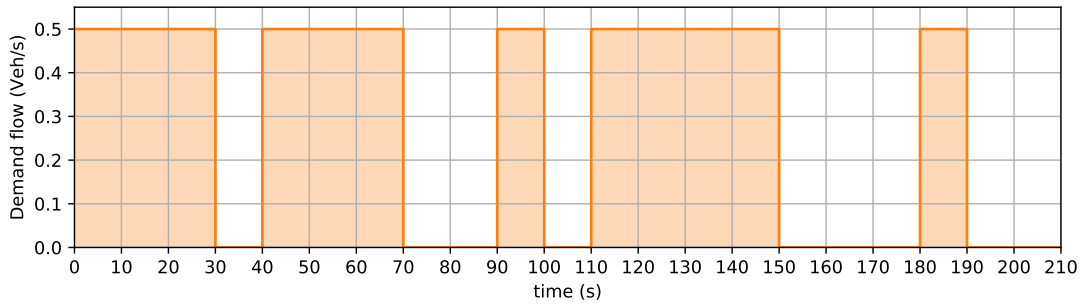


Figure 6.3: One example of a demand profile used in the experiments, with 5 platoons of varying length and arrival times as applied to an input link over 200 s. The platoons consist of 15, 15, 5, 20 and 5 vehicles respectively, with a flow rate of 0.5 vehicles/s and a density of 0.036 vehicles/m (the peak of the fundamental diagram in Fig. 6.2(a)).

CTM

To model the networks with CTM, each link must be split into cells, where the length of each cell is given by $L = v_0 \Delta t = 139$ m. This results in 1, 2 or 3 cells per link corresponding to the possible link lengths of 139, 278 or 417 m, respectively. The maximum cell occupancy is then $N = \rho^{\text{jam}} L = 20$ vehicles per cell, and the maximum flow between cells is $Q = \Delta t \text{F} = 5$ vehicles per time step. The backwards wave ratio used in constraint (C4), is given by $\frac{w}{v} = \frac{4.63}{13.89} = \frac{1}{3}$, and $I_{i,t}$ used in constraint (C8) is determined from the demand profiles. In order to evaluate the impact of withholding on the solution, we use two different values of β for the objective function weight in O2, $\beta = 0.001$ and $\beta = 0$.

LTM

The variable indexing of the LTM formulation requires knowing the integer number of time intervals for the forwards and backwards waves to travel the length of the link, giving $w_i = \frac{L}{w\Delta t} \in \{2, 4, 6\}$ and $v_i = \frac{L}{v\Delta t} \in \{1, 2, 3\}$, where $L \in \{139, 278, 417\}$ is the link length. The maximum flow rate per link is determined by the peak of the fundamental diagram in Fig. 6.2(a) and $\Delta t = 10$ s as $\text{F}_{i,j} = 5$ vehicles per time step. The cost used in determining receiving flow in constraints (L1) and (L2) is $c_i = L\rho^{\text{jam}} \in \{20, 40, 60\}$ depending on the link length L of 139, 278 or 417 m respectively.

VT

Solving the VT formulation requires construction of the lopsided network of wave paths using the set of ordinary paths $\mathcal{L}_o$ and signal paths $\mathcal{L}_s$. With the fundamental diagram in Fig. 6.2(a) we have three wave paths to consider: forward wave paths with slope $v = 13.89$, backwards wave paths with slope $w = 4.63$, and signal paths with slope 0. For the ordinary wave paths $(i, j) \in \mathcal{L}_o$ this gives a $\Delta x_{i,j} = L$ where L is the link length of 139, 278 or 417 m, and $\Delta t_{i,j} = vL \in \{10, 20, 30\}$ for the forward wave paths and $\Delta t_{i,j} = wL \in \{30, 60, 90\}$ for the backwards wave paths, corresponding to the link lengths respectively. The cost of ordinary paths is determined from the cost function in Fig. 6.2(b) as $c_{i,j} = 0$ for the forward paths, and

$c_{i,j} = 0.66\frac{L}{w} \in \{20, 40, 60\}$ for the backwards paths. For signal paths $(i, j) \in \mathcal{L}_s$, $\Delta x_{i,j} = 0$, $\Delta t_{i,j} = 10$ and the cost $c_{i,j} = 0.5\Delta t_{i,j} p_{i,j} = 5p_{i,j}$ where $p_{i,j} \in \{0, 1\}$ is the binary variable representing the signal state associated with signal path (i, j).

Microsimulation Parameters

We use IDM as a micro-simulator, and choose parameters similar to those suggested in Treiber et al. [2000], that give realistic values for urban traffic with a flow capacity of 0.5 vehicles/s and a jam density of 0.15 vehicles/m. To simulate the average conditions, we give all vehicles the same parameter values: length $l = 4.67$ m, desired velocity $v_0 = 15$ m/s, safe time headway $T = 1$ s, maximum acceleration $a = 2$ m/s^2, desired deceleration $b = 3$ m/s^2, acceleration exponent $\delta = 4$, and jam distances $s_0 = 2$ m, and $s_1 = 2$ m.

6.2.5 Evaluation

We generate the 9 different sized grids in 4 scenarios:

- one-way streets with fixed intersection spacing
- one-way streets with varying intersection spacing
- two-way streets with fixed intersection spacing
- two-way streets with varying intersection spacing

Additionally, for each of the 9 sizes, we generate 10 demand profile scenarios, where each input receives a unique demand profile of varying platoon lengths and arrival times. The 10 demand scenarios are reused in each of the 4 spacing and direction scenarios.

Experiments without Lost Time

We than take each model, and for each of the 9 network sizes and 4 scenarios, and for each of the three signal timing models, and each of the 10 demand profiles, we solve a MILP with a planning horizon of 600 s, sufficient to allow all the traffic to clear the network. We use Gurobi as the solver, running in a single thread, and record the CPU time taken to solve the MILP to within a MIP gap of 0.01% of the objective value. To evaluate the quality of the solution, first we calculate the total travel time of the solution by summing the areas under the departure curves for each output link. Then, to correct any traffic withholding, we recalculate the traffic flow using the solution policy with the LTM model for VT, and with $\beta = 0.001$ for the CTM and QTM experiments with $\beta = 0$. Using the corrected solution, we take the cumulative departure and arrival curves for each link and calculate the travel time of each vehicle through the network. By subtracting the free flow travel time, (the time for the vehicle to travel through the network at the free flow speed, without stopping), we get the delay incurred by each vehicle.

Experiments with Lost Time

For the lost time comparison we also use two different versions of the networks. For the first evaluation, the one-way fixed spacing scenario is used with 10 demand profile scenarios. We evaluate three models using signal constraints that support lost time: QTM and CTM with QTM signal constraints, and VT with SCN signal constraints. For each model and each grid size and each of the 10 demand scenarios, we solve a MILP with a planning horizon of 600's. We using Gurobi as the solver and in order to run the experiments in a reasonable time frame, we run the solver with 20 threads across 20 CPU cores. We record the CPU time to find a solution with a MIP gap of 0.01% of the objective value, along with the delay incurred by each vehicle.

For the second evaluation, we compare VT with QTM, but relax the constraint that link lengths are multiples of the time step, and can take any value between 139m and 417m. However, construction of a valid wave path lattice for VT requires that link lengths be a common multiple of $v\Delta t$ and $w\Delta t$. Since from the fundamental diagram in Fig. 6.2(a), we have $\frac{v}{w} = 3$, link lengths for VT must be rounded to the nearest multiple of $v\Delta t$ before determining the VT parameters.

In order to evaluate the impact of link length rounding on the VT formulation, we compare VT using three different values of Δt: 10 s, 5 s, and 2 s, with QTM using $\Delta t = 10$ s. For VT the link lengths are rounded such that when $\Delta t = 10$ s, $L \in \{139, 278, 417\}$, when $\Delta t = 5$ s, $L \in \{139, 208, 278, 347, 417\}$, and $\Delta t = 2$ s, $L \in \{139, 167, 194, 222, 250, 278, 306, 333, 361, 389, 417\}$. QTM has no such limitations and $L \in [139, 417]$. For each grid size and each demand scenario, and for each model and time step, we solve a MILP with a planning horizon 600s and a lost time of 10 s, and we record the CPU time taken to solve the MILP to within a MIP gap of 0.01% of the objective value. Next we adjust the solution signal plan timing to account for the lost time, and micro-simulate the network with the adjusted signal plan, using vehicles arrivals determined from the demand scenario. Then, for each vehicle in the micro-simulation, we determine the delay incurred traversing the network, and record the values.

Fundament Diagram Experiments

To evaluate the impact of shock waves on solver performance we compare QTM with VT, and CTM using several different fundamental diagrams. For the network we use the 2×2 grid network with a spacing of $v\Delta t = 139$ m between intersections, which is the minimum possible with $\Delta t = 10$ s. For VT we use the a triangular fundamental diagram with the same wave speeds as the previous experiments but with a jam density of 0.15 Vehicles/m to more closely approximate the IDM fundamental diagram. For CTM we use a trapezoidal relation with $w = v$. For QTM, we use two different versions of the solver: QTM with the same $F_{i,j}$ and Q_i as the previous experiments, and QTM with additional backwards waves in a form that more closely approximates the IDM fundamental diagram, using the constraints (Q23) and (Q24) described in Section 5.5. The fundamental diagrams are plotted in Fig. 6.10.

6.3 Results

Fixed spacing with one-way streets

Fig. 6.4 show the result of the comparison for the fixed intersection spacing with one-way streets scenario. The upper plots show, for each traffic model and each lighting model, the solve time averaged over the 10 different demand scenarios, as a function of increasing number of intersections (increasing grid size). Overlaid on each plot are error bars showing the 95% confidence interval on the mean for each grid size. The lower plots show box plots of delay distribution for every vehicle, across all the grid sizes and demand scenarios, and for each traffic model and signal model.

The upper plots show that across all the models, solve time increases exponentially with increasing number of intersections. VT has a consistent performance advantage across all on the signal models, except under SCN with more the 9 intersections, where QTM with $\beta = 0.001$ does better. The CTM and LTM show worse performance under all signal models. QTM with $\beta = 0.001$ out performs QTM with $\beta = 0$, while CTM with $\beta = 0.001$ outperforms CTM with $\beta = 0$ above 36 intersections. The 95% confidence intervals on the mean show that, for the fixed spacing, one-way scenario, across all the models there is very little variation in solve time with variation in demand profile.

The box plots show that across all traffic models and signal models, the distribution of delay is very similar with identical mean, median, and interquartile range, and the maximum delays are also similar and all are within the measurement resolution of $\Delta t = 10$ s.

Varying spacing with one-way streets

Fig. 6.5 show the result of the comparison for the varying intersection spacing with one-way streets scenario. The varying intersection spacing should present a more difficult coordination problem for the solver to optimize, and across all the traffic models and signal models, the solve time increases when compared to the fixed spacing scenario. Additionally, the 95% confidence intervals on the mean show increased variation in solve time. Overall, the trend is similar to the fixed spacing scenario, with VT performing better than the other traffic models, except under SCN where QTM does better with larger numbers of intersections. Similarly QTM with $\beta = 0.001$ outperforms QTM with $\beta = 0$, while for CTM, $\beta = 0.001$ does better then $\beta = 0$ with larger networks.

The box plots show very similar results to the fixed spacing scenario with identical mean, median and interquartile range, and the maximum delay is also similar within the resolution of $\Delta t = 10$ s.

Fixed spacing with two-way streets

Fig. 6.6 show the result of the comparison for the fixed intersection spacing with two-way streets scenario. The upper plot shows median solve time taken across all the demand scenarios with error bars indicating the 95% confidence interval in the median. Two-way streets should present a more difficult coordination problem for the solver to optimize, and, compared to the one-way scenarios, there is an across the board further increase in solve time. The 95% confidence intervals also show an increased variation in solve time.

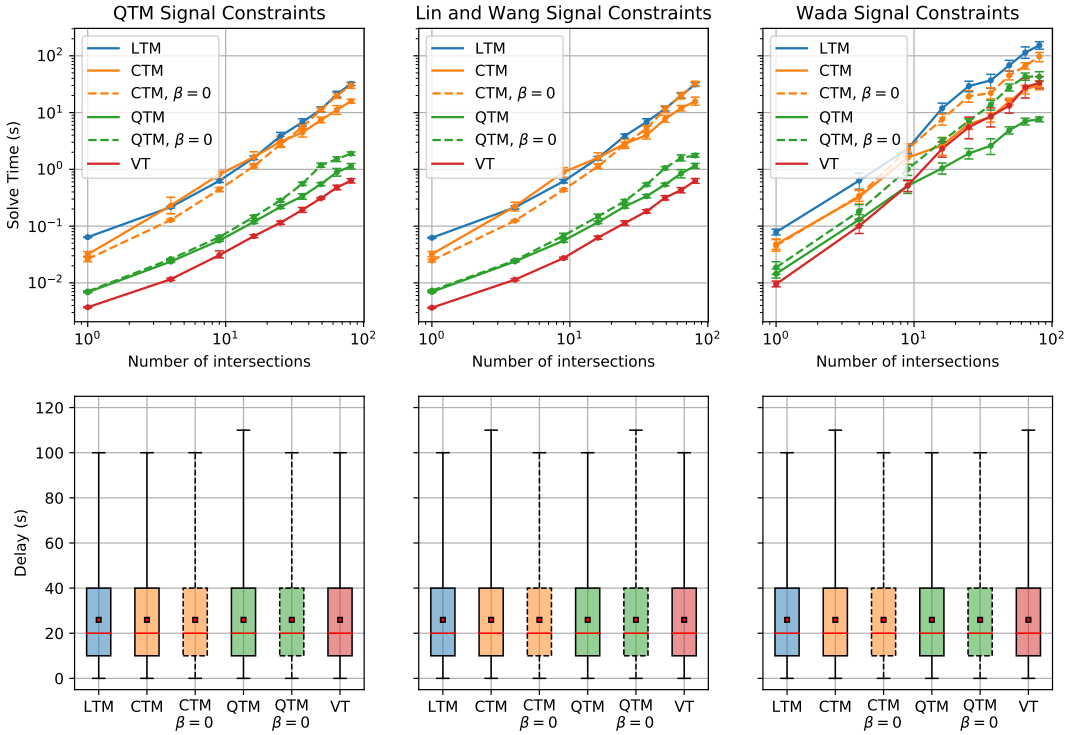


Figure 6.4: Results for the fixed intersection spacing, with one-way streets scenarios. The upper plots show solve time as a function of number of intersections, with VT showing an advantage except under SCN, and generally models with $\beta = 0.001$ do better than $\beta = 0$ with larger networks. The lower plots show the distribution of vehicle delay across all the scenarios. All the combinations of traffic model and signal model show very similar delay distribution.

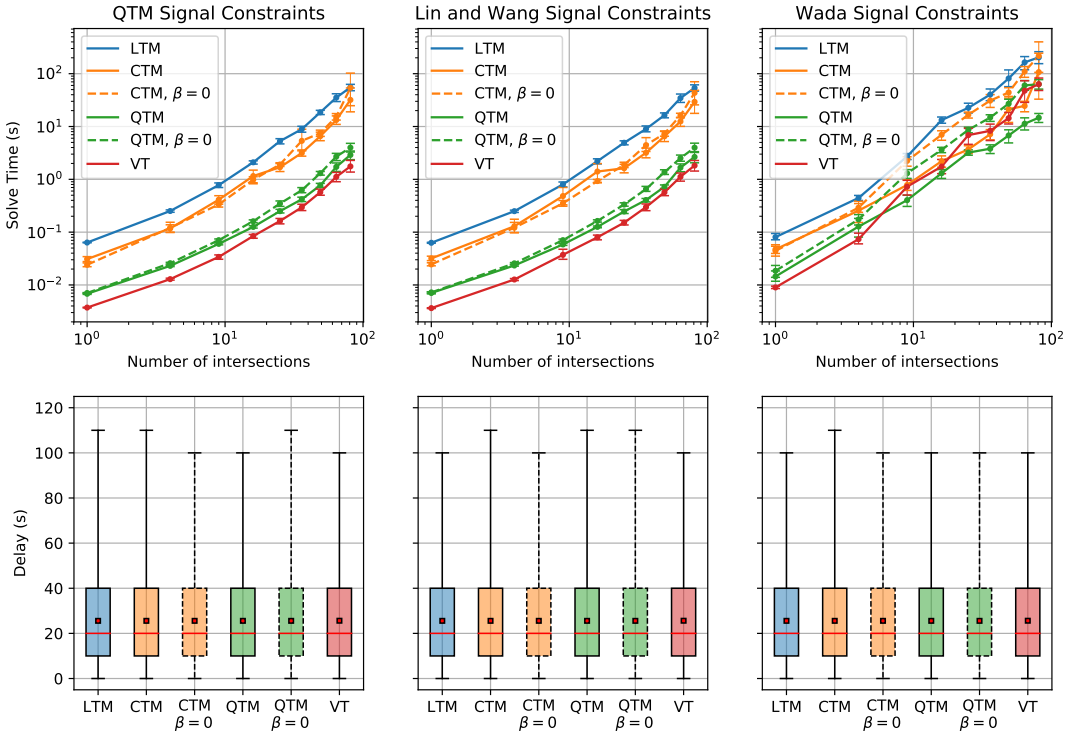


Figure 6.5: Results for the varying intersection spacing, with one-way scenarios. Solve times increasing slightly compared to the fixed spacing scenario, and the 95% confidence intervals show increased variation in solve time. Overall, the trend is similar to the fixed spacing scenario.

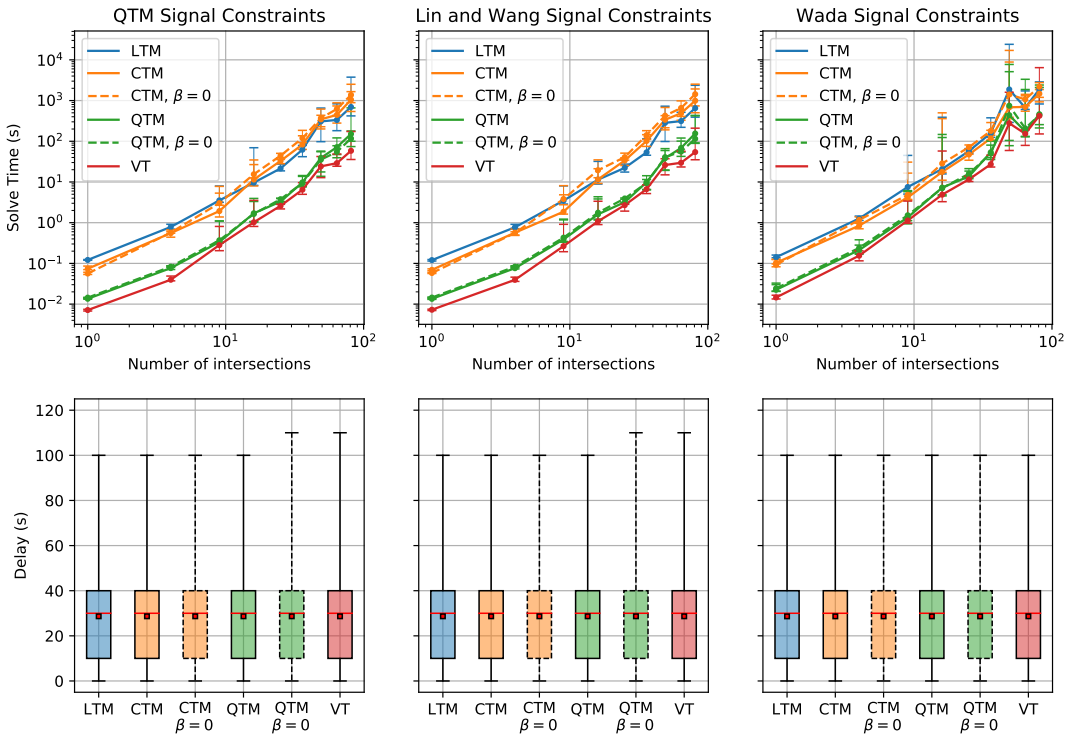


Figure 6.6: Results for the fixed intersection spacing, with two-way scenarios. Solve times increasing compared to the one-way spacing scenario, and the 95% confidence intervals show increased variation in solve time. QTM and VT consistently outperform CTM and LTM, with VT showing increased variation in solve time compared to QTM. $\beta = 0.001$ consistently outperforms $\beta = 0$.

Comparing the performance of the models, VT and QTM consistently outperform CTM and LTM, with VT showing better median performance across the demand scenarios, but with increased variation in solve time.

Additionally, we see that the QTM and CTM with $\beta = 0.001$, does better than $\beta = 0$, especially with larger networks.

The box plots show that across all traffic models and signal models, the distribution of delay is very similar with identical mean, median, and interquartile range, and the maximum delays are also similar and all are within the measurement resolution of $\Delta t = 10$ s. Compared with the one-way scenarios, there is an increase in mean and median and delay, while the interquartile range remains the same, and the maximum delay is similar within the measurement resolution of $\Delta t = 10$ s.

Varying spacing with two-way streets

Fig. 6.7 show the result of the comparison for the varying intersection spacing with two-way streets scenario. The upper plot shows median solve time taken across all the demand scenarios with error bars indicating the 95% confidence interval in the median. Compared to

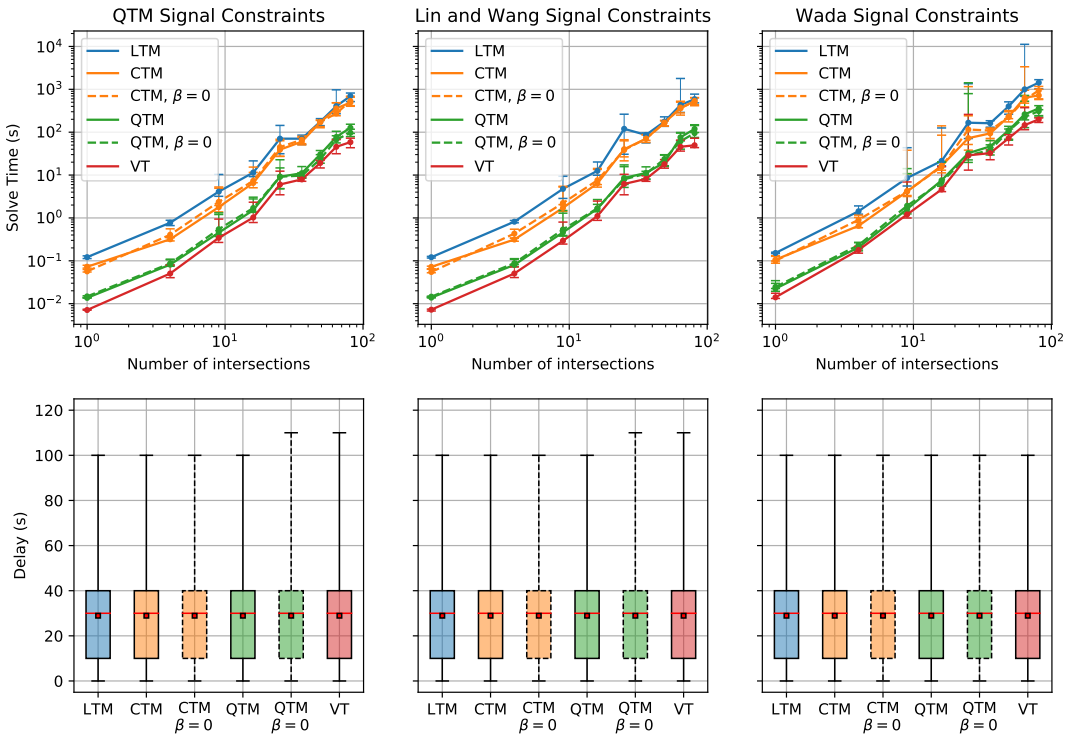


Figure 6.7: Results for the varying intersection spacing with two-way streets scenario. Compared to the fixed spacing, two-way scenario, the median solve times and confidence intervals show similar results. Similarly, the box plots show identical mean, median, interquartile range, and maximum delays.

the fixed spacing, two-way scenario, the median solve time and 95% confidence intervals also show similar results. QTM and CTM with $\beta = 0.001$, continue to do better than $\beta = 0$.

The box plots show identical interquartile range, and similar mean and median to the two-way, fixed spacing scenarios, and the maximum delays are also similar within the measurement resolution of $\Delta t = 10$ s.

Withholding and Objective Function Equivalence

An inspection of the total travel times for the results shows that all the solvers find solutions with the same total travel time with the resolution of the MIP gap of 0.01%.

Comparing the results For CTM and QTM with $\beta = 0.001$ and $\beta = 0$, we see that while there is some variation on solve time, the box plots show that the solutions are very similar within the measurement resolution. Further, the values of the total travel time are the same within the resolution of the MIP gap for both values of β and when compared with VT and LTM, showing that any traffic withholding in the solution does not impact on its quality, and further that the objective functions are equivalent.

The differences in the QTM results between $\beta = 0.001$ and $\beta = 0$ suggest that the β weights in the objective function help to guide the solver to the optimum solution. To under-

stand this, consider that when $\beta = 0$, during solving the incumbent solution can incur vehicle withholding at the exit due to the variable cost function of the equivalent shortest path problem which only yields a valid solution once the bounds have converged. This leads the solver to explore more suboptimal branches, than when $\beta = 0.001$. This is further illustrated by the increase in solve time for VT under SCN compared to QTM with $\beta = 0.001$, which has the additional feasible path problem with variable constraints in the signal network. For CTM in the one-way scenarios, due to the additional terms in objective function from the cells where $\beta = 0.001$, the effect of withholding on the incumbent solution only becomes apparent with larger networks.

Lost time comparison

As was shown in Chapter 4, modelling the lost time associated with signal changes is necessary to achieving realizable signal plans. Fig. 6.8 shows the result of solving the one-way, fixed intersection spacing scenario, over all the demand profiles, but with the addition of lost time constraints as supported by the CTM, QTM and VT models. The left plot shows the mean solve time as a function of the number of intersections, with error bars showing the 95% confidence interval on the mean. The box plots on the right show the distribution of microsimulation delay over all the demand scenarios, with all models finding similar policies, with the same mean, median, interquartile range and maximum delay within the measurement resolution of $\Delta t = 10$ s.

The result shows that both VT and QTM do better than CTM, with VT improving over QTM with increasing number of intersections. Compared with the results without lost time in Fig. 6.4, the solve times when including lost time constraints are increased for the same size network, indicating that finding an optimum solution while considering the time lost to signal changes is a more difficult problem.

Micro-simulation with varying link lengths

Accurate modelling of distances between intersections is critical for good signal coordination, and it is interesting to see the trade off between time step resolution, and signal plan quality. Additionally, increasing the number of time steps can increase the time needed to find the optimum solution, so a balance must be found. Fig. 6.9 shows the result of comparing VT using several different values of Δt with QTM using $\Delta t = 10$ s, micro-simulated on networks with varying distances between intersections. The right plot shows the mean solve times across the demand scenarios, as a function of number of intersections, with error bars giving the 95% confidence interval. As the time step size is reduced for VT, the solve time increases for all grid sizes. The right plot shows for each model and time step size, box plots of the distribution of vehicle delay in the micro-simulation. VT policies show improving delay distribution with decreasing time step, however QTM with $\Delta t = 10$ s is able to find better policies with lower maximum, median and average delay, while solving at all grid sizes in less time than VT with $\Delta t = 2$ s.

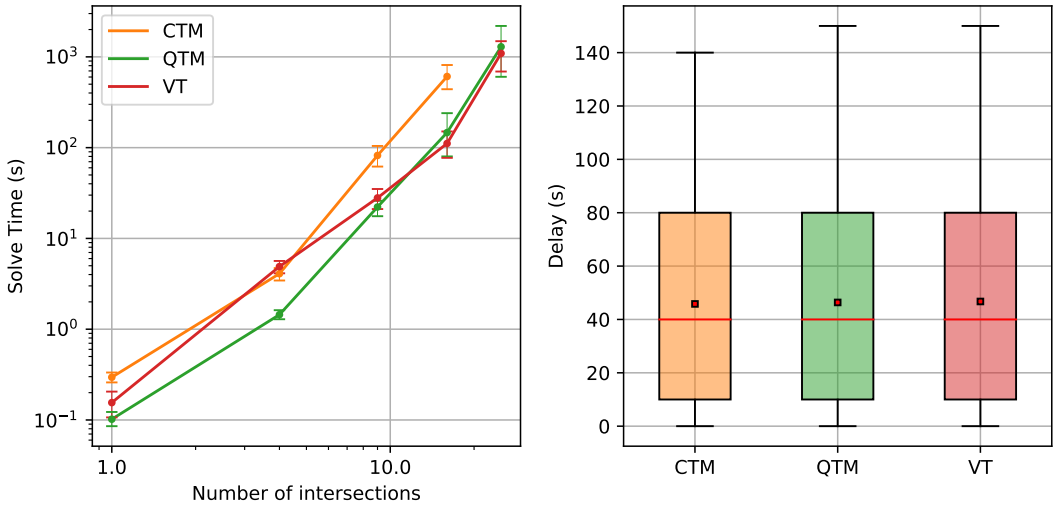


Figure 6.8: Results for the fixed intersection spacing scenario, with one-way streets and lost time constraints. Compared with the fixed intersection spacing, one-way scenario, without lost time in Fig. 6.4, the solve times are significantly increased with the addition of lost time constraints. VT using the SCN performs better than QTM and CTM with lower solver times and tighter 95% confidence intervals on the mean. The box plots show all the models find similar policies in terms of predicted delay.

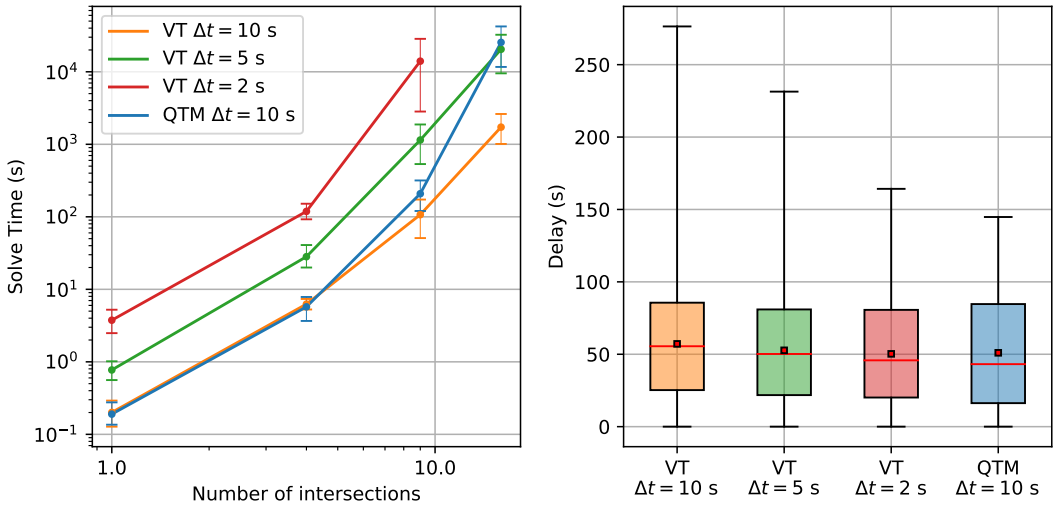


Figure 6.9: Micro-simulation comparison of QTM and VT with varying intersection spacing. The left plot shows the mean solve times with different values of Δt, and error bars give the 95% confidence interval. The right box plots show the distribution of per vehicle delay under micro-simulation of the policies, across all simulation runs. VT policies show improving delay distribution with decreasing time step, however QTM with $\Delta t = 10$ s is able to find policies with lower maximum, median and average delay, in less time than VT with $\Delta t = 2$ s.

Micro-simulation with varying Fundamental Diagrams

Fig. 6.10 shows the impact of different fundamental diagrams on performance. The lower left plot shows the three fundamental diagrams used in the experiment compared with that of the IDM microsimulator. For QTM the fundamental diagram is shown for the case where the link length is the minimum of 131 m with $\Delta t = 10$ s, and forms the same trapezoidal shape used for CTM. VT uses a triangular shape where the wave speeds are limited to be integer ratios of link length to the time step. QTM with wave extensions (QTMW) is solved using a fundamental diagram with additional wave speeds to more closely approximate the continuous relationship of the IDM fundamental diagram. The upper left plot shows the solve time with increasing traffic demand, while the upper right plot compares the increase in average delay per vehicle with increasing traffic demand, and the lower right box plots show the distribution of delay for three different levels of demand. QTM performs well with the lowest solve times and delay distribution similar to CTM, while both outperform VT in solve time, and median and average delay. QTMW outperforms the other solvers with lower mean, median, interquartile range and maximum delay, but at the cost of additional solve time compared to QTM and CTM.

Figure 6.11 shows the solutions of VT and QTMW along a street under the same demand scenario. The density of the LWR flow is calculated from each solutions cumulative arrival and departure curves and fundamental diagram, and plotted in shades of green. The IDM microsimulation of each solution's policy is overlaid with a each vehicle trajectory shown as a black trace. The red and yellow signal states over time are plotted horizontally for each intersection along the street. In Fig. 6.11(a), we can see that VT with a triangular fundamental diagram overestimates the spillback from the traffic signals, resulting in a suboptimal policy that does not align well with the microsimulation. Whereas in Fig. 6.11(b), we see that QTM with the additional backwards waves, closely predicts the vehicle flow and finds a more optimal policy.

6.4 Summary

In this chapter, we ran a suite of experiments to compare QTM with LTM, CTM and VT MILP formulations for traffic signal optimization, across a range of network topologies and demand scenarios, and using different fundamental diagrams. The results were evaluated in terms of solve time and delay distribution as a measure of quality. Overall, the results showed that when comparing both the solve time and the quality, QTM is able to find better policies with lower delay, and in less time than the other formulations.

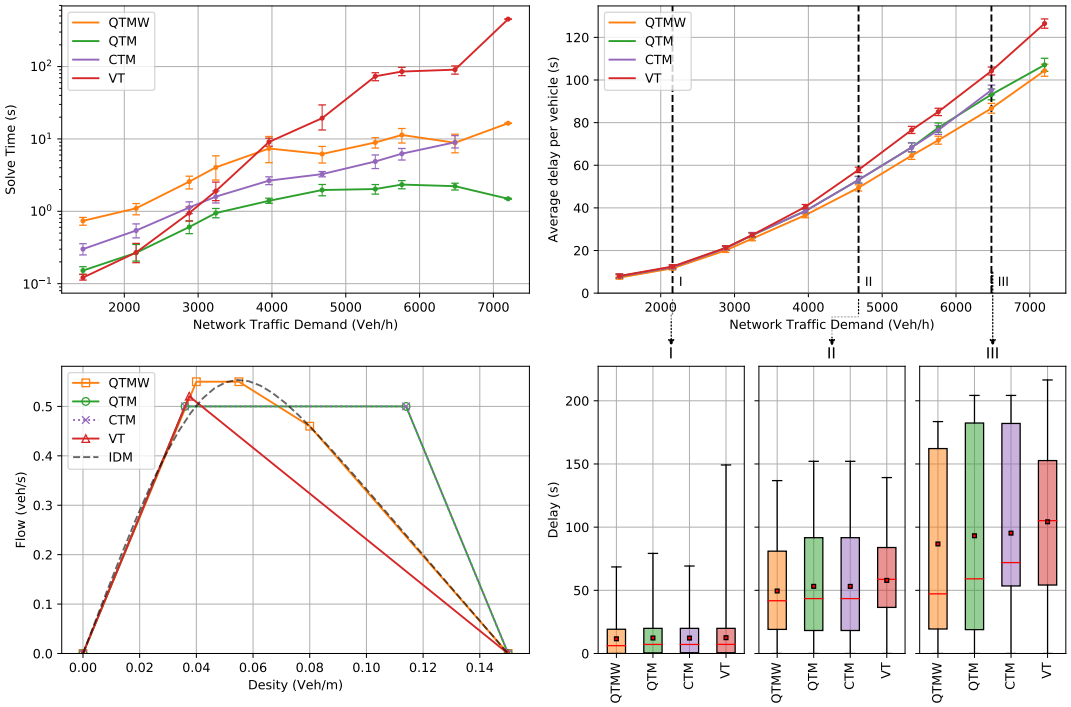
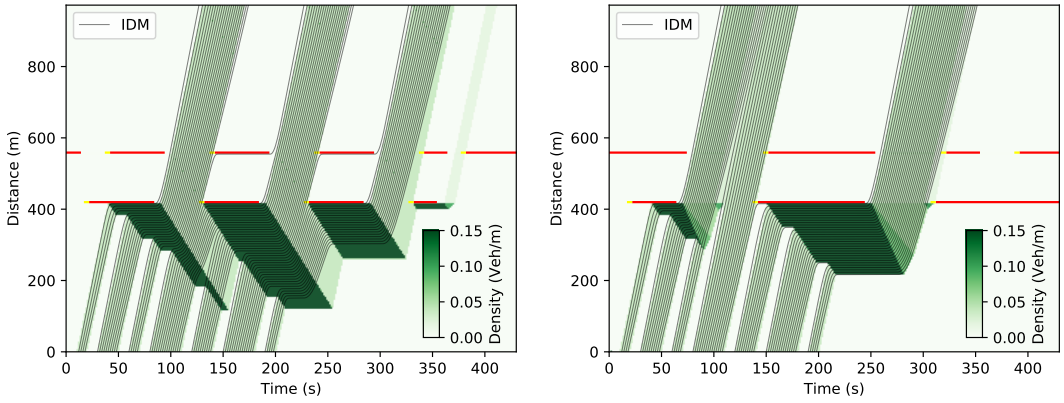


Figure 6.10: Micro-simulation comparison of fundamental diagrams with increasing traffic demand. Lower left plot shows the three fundamental diagrams compared to the IDM micro-simulator. Upper left plot shows the solve time with increasing traffic demand. Upper right plot compares the increase in average delay with increasing traffic demand, and the lower right box plots show the distribution of delay for three different levels of demand. QTM with the additional backwards waves (QTMW) shows better average delay with increasing demand, as well as improved delay distribution.



(a) (b)

Figure 6.11: An example of a VT solution compared to QTMW for the same demand scenario at 5400 Veh/h in Fig. 6.10. The density of the LWR flow predicted by each model is shaded in green and overlaid with the vehicle traces of an IDM microsimulation of the policies. (a) VT with a triangular fundamental diagram overestimates the spillback from the traffic signals, resulting in a suboptimal policy that does not align well with the micro-simulation. (b) QTMW using a more detailed fundamental diagram, closely predicts the vehicle flow and finds a more optimal policy.

Chapter 7

Conclusion

7.1 Summary

In Chapter 3, the Queue Transmission Model was introduced as a MILP model blending elements of cell based and link based models, and with the unique property of non-homogeneous time steps. It was shown how the non-homogeneous time property can be used to extend the planning horizon of a controller, without increasing the number of time intervals compared with a homogeneous time controller, such that the controller has both high fidelity in the near term and can adapt to long term predicted changes in traffic flow. Then a series of experiments using a receding horizon controller, with both non-homogenous and homogenous time steps, were used to optimize traffic signals on several different network topologies. It was demonstrated that QTM as a non-homogeneous time controller was able to find policies with substantially lower delay using less time intervals when compared with QTM as a homogeneous time controller.

Several extensions to QTM were presented in Chapter 4, along with micro-simulation validation of QTM generated signal plans. In the first extension, QTM was augmented to model the lost time associated with signal changes. By noting that the active periods of QTM signal phases represent the effective green time – the time during which traffic is actually flowing in the intersection, then lost time can be modelled by introducing additional "all red" phases in-between. After solving with lost time constraints, the start and end of the signal phases are adjusted to include the lost time in a post processing step. A comparison of micro-simulated traffic flows with QTM, showed that modelling lost time is critical to finding optimized signal plans.

Since different forms of fixed time control are still used extensively in the real world, QTM was also extended to find optimal fixed time controllers. This was done by adding constraints on the phase durations such that they can be optimized along with the offset, but remain fixed across all cycles in the signal plan. When considered as an offline process, plans generated using QTM optimized fixed time control can be deployed immediately to improve existing infrastructure.

Another real world problem studied in this thesis is how the introduction of light rail systems that share intersections with vehicle traffic impacts congestion and what can be done to minimize this impact (Chapter 4). To answer these questions, QTM was extended to include constraints for light rail schedules, such that signal plans are optimized to consider light rail crossings. The performance of these extensions was evaluated using both fixed time and optimized adaptive controllers, on several networks topologies. Various scenarios without light rail and two different light rail schedules were considered, and delay was measured across a range of traffic demand levels, using multiple different arrival patterns. The results showed that in some scenarios, when compared with the use of fixed time control beforehand, QTM optimized signals plans can effectively nullify the impact of introducing light rail.

In Chapter 5 we used Variational Theory to show that QTM, CTM, LTM and VT MILPs are equivalent to solving a shortest path problem representing the discrete Hamilton Jacobi form of the LWR kinematic wave equation. We then showed that several commonly used objective functions to minimize delay, all equivalently minimize the total travel time of the network by maximizing the area under the total cumulative departure curve. The VT equivalence, along with the objective function equivalence, demonstrates that:

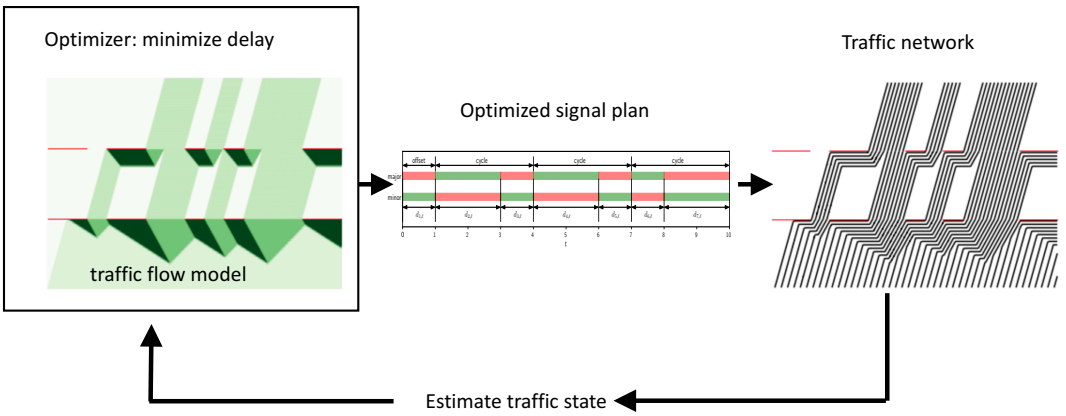


Figure 7.1: A closed loop model predictive controller for traffic signal optimization, must estimate the state of the traffic network from sensor data, and learn the model parameters. Uncertainty in the state estimate must then be considered by the optimizer.

- QTM models an LWR traffic flow using a trapezoidal fundamental diagram.
- Removing the objective function terms to prevent traffic withholding in CTM and QTM does not impact on the optimality of the policies, since withholding within any time interval implies that the equivalent node in the VT lattice network is not on any shortest wave path from a boundary node to an exit node.

We then further extended QTM with multiple shockwaves to model any concave fundamental diagram with arbitrary precision, and showed using microsimulation that it can be used to further improve the quality of the solutions.

We then derived a continuous time MILP for solving QTM. This formulation has the advantage of requiring only the minimum number of intervals for a QTM solution, with time steps only at change points in the queue function. This gives the exact solution to a QTM traffic flow through a signalized network. We then developed an objective function to minimize delay and find an optimal continuous time signal plan.

Finally, in Chapter 6, we evaluated the performance of QTM compared with CTM, LTM and VT MILP formulations for global signal optimization. A set of controllers with different signal constraint models was evaluated on a series of grid shaped traffic networks, ranging in size up to 81 intersections, with multiple demand scenarios. Using multiple topologies of one-way and two-way streets, fixed intersection spacing, discretely varying intersection spacing, and continuously varying intersection spacing, we showed that QTM is able to outperform the other formulations, when considering both the solve time and the quality of the solution under micro-simulation.

7.2 Future Work

This thesis has focused on MILP based traffic signal optimization models, but a closed loop model predictive controller must also estimate the traffic state and learn the model parameters

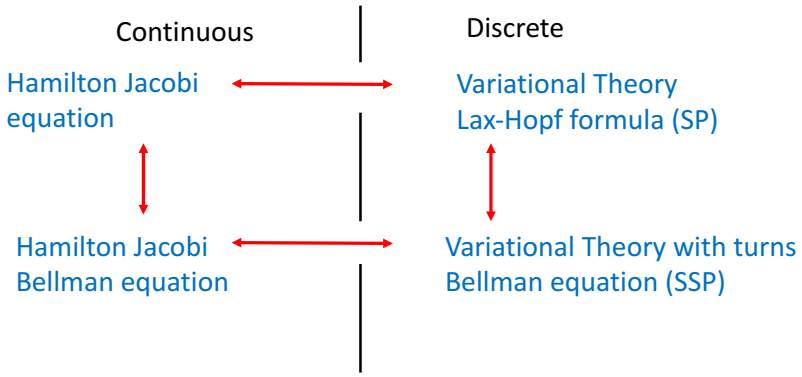


Figure 7.2: Relationship between the Hamilton Jacobi equation and the Bellman equation. Variational Theory (VT) can solve a discrete form of the Hamilton Jacobi equation, using the Lax-Hopf formula, as a shortest path problem (SP). With the addition of turn probabilities, VT takes the form of a stochastic shortest path problem (SSP), equivalent to solving a Markov Decision Problem value function using the Bellman equation.

as illustrated in Fig. 7.1. This presents additional challenges:

- Irregular sensor coverage and noise results in partial observability of the traffic state. In turn, this leads to uncertainty in estimating the model parameters.
- This uncertainty should then be considered by the optimizer.

Scenarios involving turning movements were not evaluated in this work and are an additional source of uncertainty. Indeed, in Chapter 5 we derived an extension for VT using turn probabilities, and saw that it is a stochastic shortest path problem, equivalent to maximising a Markov Decision Process (MDP) value function [Littman et al., 2013] (See Fig. 7.2). When considered along with the partial observability of the state, the result is a partially observable MDP (POMDP) which is known to be intractable to solve optimally. However, recently there have been promising results in approximately solving large POMDP's using Monte Carlo Tree search [Silver and Veness, 2010; Kurniawati and Yadav, 2016; Sunberg and Kochenderfer, 2018], and future work would be to apply these frameworks to traffic signal optimization, and investigate the benefits when compared to a MILP based approach.

While this thesis has demonstrated improved scalability of MILP's for global traffic signal optimization, the results in Chapters 4 and 6 show there are practical limits. Future work would be to explore methods of decomposition and partitioning in order to improve scalability even further.

Objective functions for traffic signal optimization have received little focus in the literature. Methods have been presented for minimizing total delay [Lo, 1998], number of stops [cite lin and wang], emission reduction [Han et al., 2016a], and bounding shockwaves [Han et al., 2012], but none have looked at the distribution of delay. An Important future work to address this deficiency will need to consider fairness measures beyond simple averages or upper bounds on delay and, for instance, consider constraints on the maximum observed delay by the top 5-percentile vehicles

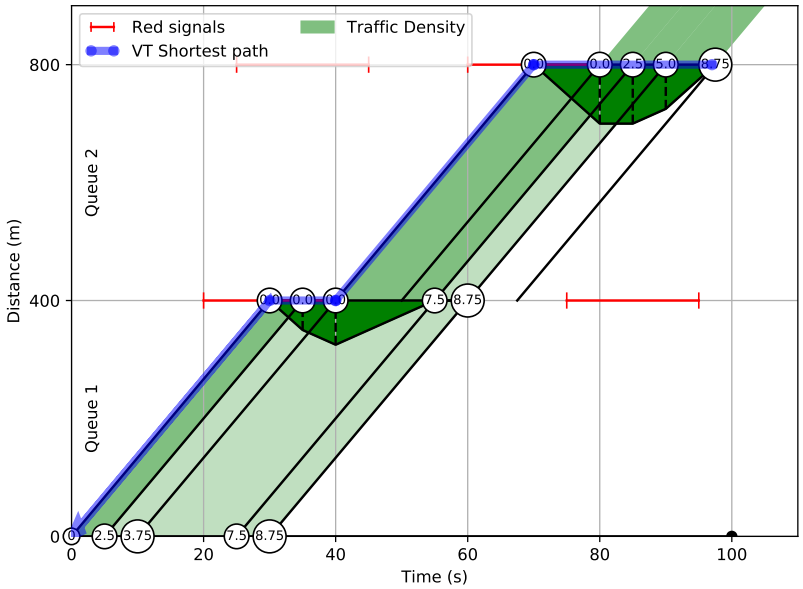


Figure 7.3: Example of the QTM continuous time solution from Section 5.7.2, along with the equivalent VT shortest path solution. Cumulative vehicle counts are labelled at the ends of each time interval. The VT solution suggests that the number of intervals needed to find the cumulative departure values is less than the four per queue needed for the QTM continuous time solution.

Performance of the QTM continuous time MILP needs to be evaluated, with the possibility of a hybrid formulation that uses continuous time for the short-term control horizon and discrete non-homogenous QTM to extend to the planning horizon. Figure 7.3 shows the result when a shortest path solution is overlaid on the QTM continuous time solution from the example in Section 5.7.2, it is clear that there is room for further reduction in the number of time intervals needed to find the solution. Future work would be to extend the fixed wave path lattice of VT with the QTM continuous framework, and solve it as a dynamic shortest path problem.

Appendix

A.1 Code Repository

The code for all the experiments in this thesis can be found in the GitHub repository:
https://github.com/iainguilliard/qtm

Bibliography

Alecsandru, C.; Quddus, A.; Huang, K. C.; Rouhieh, B.; Khan, A. R.; and Zeng, Q., 2011. An assessment of the cell-transmission traffic flow paradigm: Development and applications. In *Transportation Research Board 90th Annual Meeting*, 11-1152.

Beard, C. and Ziliaskopoulos, A., 2006. System optimal signal optimization formulation. *Transportation research record*, 1978, 1 (2006), 102–112.

Christofa, E.; Ampountolas, K.; and Skabardonis, A., 2016. Arterial traffic signal optimization: A person-based approach. *Transportation Research Part C: Emerging Technologies*, 66 (2016), 27–47.

Daganzo, C. F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological*, 28, 4 (1994), 269–287.

Daganzo, C. F., 1995. The cell transmission model, part ii: network traffic. *Transportation Research Part B: Methodological*, 29, 2 (1995), 79–93.

Daganzo, C. F., 2005a. A variational formulation of kinematic waves: basic theory and complex boundary conditions. *Transportation Research Part B: Methodological*, 39, 2 (2005), 187–196.

Daganzo, C. F., 2005b. A variational formulation of kinematic waves: Solution methods. *Transportation Research Part B: Methodological*, 39, 10 (2005), 934–950.

El-Tantawy, S.; Abdulhai, B.; and Abdelgawad, H., 2013. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atsc): methodology and large-scale application on downtown Toronto. *Intelligent Transportation Systems, IEEE Transactions on*, 14, 3 (2013), 1140–1150.

Garcia, C. E.; Prett, D. M.; and Morari, M., 1989. Model predictive control: theory and practice—a survey. *Automatica*, 25, 3 (1989), 335–348.

Gartner, N.; Little, J. D.; and Gabbay, H., 1974. Optimization of traffic signal settings in networks by mixed-integer linear programming. Technical report, DTIC Document.

Gartner, N. H. and Stamatiadis, C., 2002. Arterial-based control of traffic flow in urban grid networks. *Mathematical and computer modelling*, 35, 5 (2002), 657–671.

Gomes, G. and Horowitz, R., 2006. Optimal freeway ramp metering using the asymmetric cell transmission model. *Transportation Research Part C: Emerging Technologies*, 14, 4 (2006), 244–262.

Guilliard, I.; Sanner, S.; Trevizan, F. W.; and Williams, B. C., 2016. Nonhomogeneous time mixed integer linear programming formulation for traffic signal control. *Transportation Research Record: Journal of the Transportation Research Board*, 2595 (2016), 128–138. doi: 10.3141/2595-14.

Guilliard, I.; Trevizan, F. W.; and Sanner, S., 2020. Mitigating the impact of light rail on urban traffic networks using mixed integer linear programming. *IET Intelligent Transport Systems*, 14, 6 (2020), 523–533. doi: 10.1049/2019.0277. URL https://doi.org/10.1049/iet-its.2019.0277.

Hajiahmadi, M.; De Schutter, B.; and Hellendoorn, H., 2012. Model predictive traffic control: A mixed-logical dynamic approach based on the link transmission model. *IFAC Proceedings Volumes*, 45, 24 (2012), 144–149.

Han, K.; Friesz, T. L.; and Yao, T., 2012. A link-based mixed integer lp approach for adaptive traffic signal control. *arXiv preprint arXiv:1211.4625*, (2012).

Han, K.; Liu, H.; Gayah, V. V.; Friesz, T. L.; and Yao, T., 2016a. A robust optimization approach for dynamic traffic signal control with emission considerations. *Transportation Research Part C: Emerging Technologies*, 70 (2016), 3–26.

Han, K.; Piccoli, B.; and Szeto, W., 2016b. Continuous-time link-based kinematic wave model: formulation, solution existence, and well-posedness. *Transportmetrica B: Transport Dynamics*, 4, 3 (2016), 187–222.

He, Q.; Head, K. L.; and Ding, J., 2011. Pamscod: Platoon-based arterial multi-modal signal control with online data. *Procedia-Social and Behavioral Sciences*, 17 (2011), 462–489.

He, Q.; Head, K. L.; and Ding, J., 2014. Multi-modal traffic signal control with priority, signal actuation and coordination. *Transportation Research Part C: Emerging Technologies*, 46 (2014), 65–82.

He, Q.; Lin, W.-H.; Liu, H.; and Head, K. L., 2010. Heuristic algorithms to solve 0–1 mixed integer lp formulations for traffic signal control problems. In *Service Operations and Logistics and Informatics (SOLI), 2010 IEEE International Conference on*, 118–124. IEEE.

Huang, K. C., 2011. *Traffic Simulation Model for Urban Networks: CTM-URBAN.* Ph.D. thesis, Concordia University.

Hunt, P. B.; Robertson, D. I.; Bretherton, R. D.; and Winton, R. I., 1981. SCOOT–a traffic responsive method of coordinating signals. Technical report, Transportation Road Research Lab, Crowthorne, U.K.

Islam, S. B. A.; Hajbabaie, A.; and Aziz, H. A., 2020. A real-time network-level traffic signal control methodology with partial connected vehicle information. *Transportation Research Part C: Emerging Technologies*, 121 (2020), 102830. doi: https://doi.org/10.1016/j.trc.2020.102830. URL https://www.sciencedirect.com/science/article/pii/S0968090X20307348.

Jabari, S. E. and Liu, H. X., 2012. A stochastic model of traffic flow: Theoretical foundations. *Transportation Research Part B: Methodological*, 46, 1 (2012), 156–174.

Kim, Y., 2002. Online traffic flow model applying dynamic flow-density relation. Technical report, Int. At. Energy Agency.

Kühne, R. and Gartner, N., 2011. 75 years of the fundamental diagram for traffic flow theory: Greenshields symposium. *Transportation Research Board, Washington, DC*, (2011).

Kurniawati, H. and Yadav, V., 2016. An online pomdp solver for uncertainty planning in dynamic environment. In *Robotics Research*, 611–629. Springer.

Li, J. and Zhang, H. M., 2014. Coupled linear programming approach for decentralized control of urban traffic. *Transportation Research Record*, 2439, 1 (2014), 83–93. doi: 10.3141/2439-08. URL https://doi.org/10.3141/2439-08.

Lighthill, M. and Whitham, G., 1955. On kinematic waves. ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London Series A*, 229 (1955), 317–345.

Lin, W.-H. and Wang, C., 2004. An enhanced 0-1 mixed-integer lp formulation for traffic signal control. *Intelligent Transportation Systems, IEEE Transactions on*, 5, 4 (2004), 238–245.

Littman, M. L.; Dean, T. L.; and Kaelbling, L. P., 2013. On the complexity of solving markov decision problems. *arXiv preprint arXiv:1302.4971*, (2013).

Lo, H. K., 1998. A novel traffic signal control formulation. *Transportation Research Part A: Policy and Practice*, 33, 6 (1998), 433–448.

Lo, H. K.; Chang, E.; and Chan, Y. C., 1999. Dynamic network traffic control. *Transportation Research Part A: Policy and Practice*, 35, 8 (1999), 721–744.

Lu, S.; Dai, S.; and Liu, X., 2011. A discrete traffic kinetic model–integrating the lagged cell transmission and continuous traffic kinetic models. *Transportation Research Part C: Emerging Technologies*, 19, 2 (2011), 196–205.

Muralidharan, A.; Dervisoglu, G.; and Horowitz, R., 2009. Freeway traffic flow simulation using the link node cell transmission model. In *American Control Conference, 2009. ACC'09.*, 2916–2921. IEEE.

Newell, G. F., 1993a. A simplified theory of kinematic waves in highway traffic, part i: General theory. *Transportation Research Part B: Methodological*, 27, 4 (1993), 281–287.

Newell, G. F., 1993b. A simplified theory of kinematic waves in highway traffic, part ii: Queueing at freeway bottlenecks. *Transportation Research Part B: Methodological*, 27, 4 (1993), 289–303.

Newell, G. F., 1993c. A simplified theory of kinematic waves in highway traffic, part iii: Multi-destination flows. *Transportation Research Part B: Methodological*, 27, 4 (1993), 305–313.

Richards, P. I., 1956. Shock waves on the highway. *Operations research*, 4, 1 (1956), 42–51.

Silver, D. and Veness, J., 2010. Monte-carlo planning in large pomdps. In *Advances in neural information processing systems*, 2164–2172.

Sims, A. G. and Dobinson, K. W., 1980. SCAT–The Sydney Co-ordinated Adaptive Traffic System: Philosophy and benefits. *IEEE Transactions on Vehicular Technology*, 29 (1980).

Smith, S.; Barlow, G.; Xie, X.-F.; and Rubinstein, Z., 2013. SURTRAC: Scalable Urban Traffic Control. In *Transportation Research Board 92nd Annual Meeting Compendium of Papers*. Transportation Research Board.

Stevanovic, J.; Stevanovic, A.; Martin, P. T.; and Bauer, T., 2008. Stochastic optimization of traffic control and transit priority settings in vissim. *Transportation Research Part C: Emerging Technologies*, 16, 3 (2008), 332 – 349. doi: 10.1016/2008.01.002. URL https://doi.org/10.1016/j.trc.2008.01.002. Emerging Commercial Technologies.

Sumalee, A.; Zhong, R.; Pan, T.; and Szeto, W., 2011. Stochastic cell transmission model (sctm): A stochastic dynamic traffic model for traffic state surveillance and assignment. *Transportation Research Part B: Methodological*, 45, 3 (2011), 507–533.

Sunberg, Z. N. and Kochenderfer, M. J., 2018. Online algorithms for pomdps with continuous state, action, and observation spaces. In *Twenty-Eighth International Conference on Automated Planning and Scheduling*.

Thompson, G. L., 2003. Defining an alternative future: Birth of the light rail movement in north america. In *Transportation Research Board (TRB) Annual Meeting*.

Treiber, M.; Hennecke, A.; and Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62, 2 (2000), 1805.

Varaiya, P., 2013. Max pressure control of a network of signalized intersections. *Transportation Research Part C: Emerging Technologies*, 36 (2013), 177–195.

Wada, K.; Usui, K.; Takigawa, T.; and Kuwahara, M., 2017. An optimization modeling of coordinated traffic signal control based on the variational theory and its stochastic extension. *Transportation research procedia*, 23 (2017), 624–644.

Webster, F. V., 1958. Traffic signal settings. Technical report, .

Wolshon, B. and Pande, A., 2016. *Traffic engineering handbook*. John Wiley & Sons.

Xiaojian, H.; Wei, W.; and Sheng, H., 2010. Urban traffic flow prediction with variable cell transmission model. *Journal of Transportation Systems Engineering and Information Technology*, 10, 4 (2010), 73–78.

Xie, X.-F.; Smith, S. F.; and Barlow, G. J., 2012. Schedule-driven coordination for real-time traffic network control. In *ICAPS*.

Yperman, I.; Logghe, S.; and Immers, B., 2005. The link transmission model: an efficient implementation of the kinematic wave theory in traffic networks. In *Proceedings of the 10th EWGT Meeting*, 122–127. Poznan Poland.